The best fitting of three contemporary observer models reveals how participants' strategy influences the window of subjective synchrony

Abbreviated Title: Modelling the window of subjective synchrony

Kielan Yarrow[1*], Joshua A. Solomon[2], Derek H. Arnold[3] & Warrick Roseboom[4,5]

[1] Department of Psychology, City, University of London, London EC1V 0HB
[2] Centre for applied vision research, City, University of London, London EC1V 0HB
[3] School of Psychology, The University of Queensland, Brisbane QLD 4072
[4] Department of Informatics, School of Engineering and Informatics, University of Sussex, Falmer, Brighton, BN1 9QJ
[5] School of Psychology, University of Sussex, Falmer, Brighton, BN1 9QH

* Author for correspondence:

Kielan Yarrow,
Rhind Building,
City, University of London,
Northampton Square,
London EC1V 0HB

Tel: +44 (0)20 7040 8530
Email: kielan.yarrow.1@city.ac.uk

Word count (main text excluding abstract, references & appendices): 12,212

Abstract

When experimenters vary the timing between two intersensory events, and participants judge their simultaneity, an inverse-U-shaped psychometric function is obtained. Typically, this *simultaneity function* is first fitted with a model for each participant separately, before best-fitting parameters are utilised (for example compared across conditions) in the second stage of a two-step inferential procedure. Often, simultaneity-function width is interpreted as representing sensitivity to asynchrony, and/or ascribed theoretical equivalence to a window of multisensory temporal binding. Here, we instead fit a single (principled) multilevel model to data from the entire group and across several conditions at once. By asking 20 participants to sometimes be more conservative in their judgments, we demonstrate how the width of the simultaneity function is prone to strategic change and thus questionable as a measure of either sensitivity to asynchrony or multisensory binding. By repeating our analysis with three different models (two implying a decision based directly on subjective asynchrony, and a third deriving this decision from the correlation between filtered responses to sensory inputs) we find that the first model, which hypothesises, in particular, Gaussian latency noise and difficulty maintaining the stability of decision criteria across trials, is most plausible for these data.

Keywords

Time perception, timing, simultaneity, synchrony, intersensory, Bayesian, multilevel models.

Public Significance

Psychologists have made their competing theories about how humans are able to perceive the relative timing of events concrete by formulating mathematical models that attempt to describe behaviour in specific experimental tasks. Here, we focus on one such task and show that people's reports about simultaneity are inherently subjective, as implied by several current models. We also find that the best-performing of these models explains inconsistencies when responding repeatedly to objectively identical pairs of events by positing inconsistencies in both the time it takes for neural messages to propagate through the brain, and how those messages are then interpreted to form a decision.

The best fitting of three contemporary observer models reveals how participants' strategy influences the window of subjective synchrony

The case has been made that the late eighteenth-century study of individual differences in the time at which two events appear simultaneous was actually the founding question for experimental psychology (Mollon & Perkins, 1996). Interest in this topic endures, and various tasks have been developed over the years to help probe the human sense of relative time. In one such task, known as the simultaneity judgment (SJ), participants are exposed to pairs of stimuli separated by a range of asynchronies, and must judge each such pair to be either simultaneous or not (e.g. Allan, 1975a). In the intermodal variant of this task, the two stimuli affect different senses, most typically vision and audition.

This intermodal simultaneity-judgment task has proved popular with researchers for at least three reasons. Firstly, for those whose primary interest is in understanding the mechanisms by which we perceive relative time, the intersensory task seems to require the use of a specifically *temporal* system, rather than allowing participants to fall back on alternative intramodal cues that are processed by specialist systems, such as visual motion detectors (Cass & Van der Burg, 2014). Secondly, participants seem to find the simultaneity-judgment task less onerous to perform than the most popular alternative, the temporal order judgment (TOJ) task (Love et al., 2013). Thirdly, temporal coincidence provides a powerful cue that events originating from different sensory modalities have a common cause. Hence the determination of synchrony seems a necessary step towards achieving another important cognitive operation: Multisensory integration. Indeed, such integration is often found across a limited range of sub-second physical asynchronies, supporting the concept of a temporal binding window within which multisensory integration can occur (Diederich & Colonius, 2015; Holmes & Spence, 2005; Meredith et al., 1987).

Despite its popularity, the simultaneity-judgment task presents some challenges. In particular, the data it produces are not amenable to treatment via standard models of the psychometric function, which predict monotonic and S-shaped (sigmoidal) functions as responses shift from one category of binary judgment to another (e.g. Wichmann & Hill, 2001). By contrast, psychometric functions for simultaneity judgments (hereafter termed *simultaneity functions*) first rise, then fall, as asynchronies approach and then recede from zero (skip ahead to the results for multiple examples). Researchers have addressed this problem in various ways (García-Pérez & Alcalá-Quintana, 2012a; Lee & Noppeney, 2011; Schneider & Bavelier, 2003; Stone et al., 2002; van Eijk et al., 2008; Yarrow et al., 2011) including via the application of formal observer models.

In this paper, we have two broad aims. The first is to make an initial determination regarding which current model of the simultaneity judgment shows most promise. This necessitates that we review several models. In so doing, we also provide groundwork for our second goal, which is to caution researchers against making uncritical interpretations regarding summary measures, particularly relating to the width of the simultaneity function. With these goals in mind, the remainder of the introduction will progress as follows. First, we outline recent practice with regard to the analysis of simultaneity judgments and highlight some interpretative issues. Next, we describe three models of the simultaneity judgment (García-Pérez & Alcalá-Quintana, 2012a; Parise & Ernst, 2016; Yarrow et al., 2011). We then conclude the introduction by outlining an experiment that provides a suitable data set with which to both compare models and demonstrate the dependence of the simultaneity function on strategic decisions made by the participant.

Recent treatments of simultaneity-judgment data

As noted above, data from many psychophysical tasks are routinely summarised via models that predict sigmoidal psychometric functions. This prediction is premised on the assumption that each episode exposes the participant to some continuous quantity, hereafter referred to as a decision variable, which is a monotonic transform of the sensory input. For example, a single temporal order judgment trial might yield, as a decision variable, the stimulus-onset asynchrony

(SOA) between a flash and a beep. This quantity is then classified relative to a single criterion (for example above/below zero) to form a binary judgment.

Common practice is to fit the judgments from each participant / condition with a single such sigmoidal psychometric function. The parameters of this function will then have meaning in relation to the underlying model that justifies their use – for example, the mean of a fitted cumulative Gaussian function describes the position of a hypothetical decision criterion. Parameters can be compared across conditions, or correlated with other variables, as a second (inferential) step. Alternatively, all participants and conditions can be fitted at the same time within a multilevel model (Goldstein & McDonald, 1988). Such models acknowledge the clustering of individual data points (here, responses within participants) and explicitly model random variation across clusters (here, differences between participants across the group; Moscatelli et al., 2012; Prins & Kingdom, 2018).

In the case of the simultaneity judgment, properly formulated models of the psychometric function (e.g. Schneider & Bavelier, 2003) seem not to have been widely appreciated. Principled models do exist for simultaneity judgments, and relevant authors have sometimes made model-fitting code available, at least for fits to a single participant/condition at a time (Alcalá-Quintana & García-Pérez, 2013; Yarrow et al., 2016; Yarrow, 2018). However, a tradition has emerged in which researchers (including ourselves) instead resort to fitting a descriptive function that has no basis as a model of participants' actual behaviour (for example, Roseboom & Arnold, 2011).[1]

Popular approaches for treating simultaneity-judgment data include fitting a Gaussian function (Stone et al., 2002), or the piecewise fitting of two sigmoids (van Eijk et al., 2008). While we acknowledge the appeal of recent precedent when making analytic decisions, it is difficult to recommend this tradition for future research. It is worth noting that in fitting a Gaussian to simultaneity-judgment data, researchers are not remaining agnostic about the underlying model

[1] Regrettably, and presumably for reasons of simplicity, this is sometimes done by minimising squared error. This approach does not weight data points in proportion to their true likelihoods when producing parameter estimates for models predicting binary data.

that generated the data (as per non-parametric approaches like that of Lee and Noppeney, 2011). Rather, they are committing to a model, but one which is unlikely to be correct because it is not justified by any hypothesised process. Furthermore, the parameters that are derived (for example the width of a fitted Gaussian) have no relation to hypothetical cognitive operations, such as those that are laid out in principled observer models. This may encourage interpretations based on intuition and/or supposition.

By way of example, in recent years it has become fashionable to measure “temporal binding windows” using just the simultaneity-judgment task, and interpret differences between groups or conditions as indicative of differences in the temporal sensitivity of integration processes (e.g. Chen et al., 2017; Foucher et al., 2007; Habets et al., 2017; Hillock et al., 2011; Lee & Noppeney, 2011; Marsicano et al., 2022; Navarra & Fernández-Prieto, 2020; Noel et al., 2017; Scarpina et al., 2016; Stevenson et al., 2014; Zampini et al., 2005). While of considerable interest, we believe that much of this work does not include sufficiently explicit caveats about the processes that might generate the window of simultaneity, potentially misrepresenting the underlying cause(s) of differences between conditions/groups. It seems to us that this summary measure has poor face validity to measure the conceptually distinct temporal-binding window. Hence one of our goals here is to advocate more explicit recognition of the fundamentally subjective nature of the window derived from simultaneity judgments.

Some such subjective flexibility affecting the window of subjective simultaneity is predictable, as the simultaneity-judgment task is conceptually akin to a classic detection task, where observers must decide if a weak signal (for example, a very dim light or very quiet sound) is present or not. In the detection task, it is tempting to believe that signals can be detected only when they exceed some minimal value. Signals below this hard threshold would produce the categorical internal state – “I saw nothing”. However, an alternative idea, prominent since the middle of the twentieth century, is that internal states are continuous, but decision boundaries are applied to them to generate categorical responses. This debate spawned signal detection theory, in which the tendency to declare a stimulus as present depends upon the placement of a decision criterion *c*, that is distinguishable from perceptual sensitivity limited by internal noise – *d*′ (Green & Swets, 1966; Macmillan & Creelman, 2005). It seems reasonable to assume, in line with this tradition, that the perceived extent of multisensory (a)synchrony is probably also derived from a continuous internal variable, and that categorising this internal variable to judge simultaneity is a decision process. We make this notion explicit next, by describing some plausible models of the simultaneity judgment.

Observer models of the simultaneity-judgment task

In the current work, we will consider three observer models of the simultaneity-judgment task, selected for the following reasons. Firstly, they have each seen recent use in the literature. Secondly, they each have a mechanism for explaining commonly obtained subtle asymmetries in the shape of the simultaneity function. Finally, they each include parameters that can vary in order to explain strategic changes in behaviour. To our knowledge, their goodnesses-of-fit have not previously been directly compared, allowing us to do so here for the first time. The models are schematised in Figure 1.

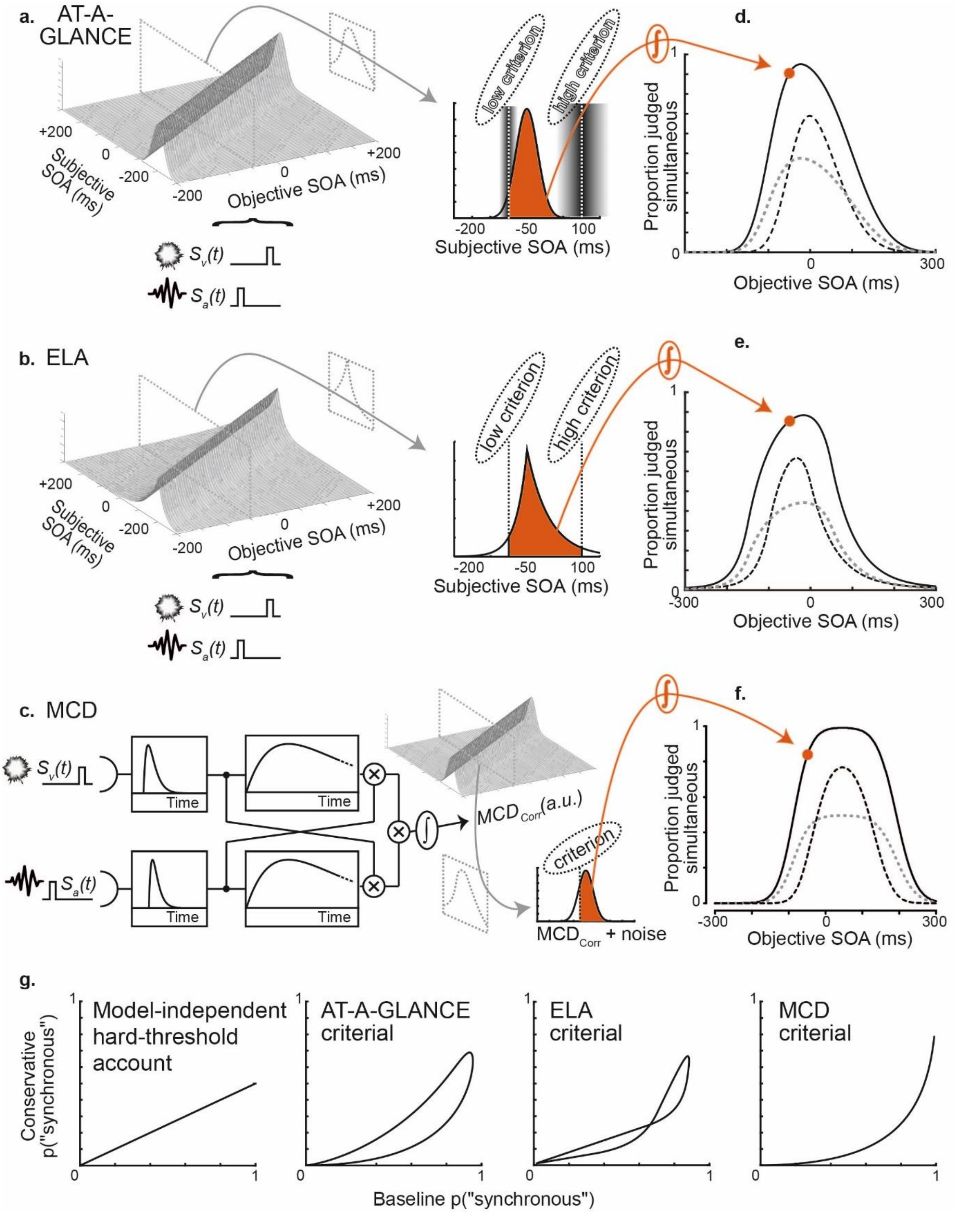

Figure 1. Schematic of models and predictions. ***(a-b)*** *In both AT-A-GLANCE and ELA models a decision centre receives both visual and auditory signals, and hence their difference in arrival times. In an experiment, each stimulus onset asynchrony (SOA) value is presented many times, yielding a noisy distribution of internal responses (subjective SOAs). The resulting probability density function (PDF) is*

shown for the example of a -50 ms SOA. Participants judge the trial as synchronous when the subjective SOA falls between two decision criteria (solid red region). For the AT-A-GLANCE model only, variable shading around the criteria indicates additional criterion noise; each criterion is most likely to be placed where the shading is darkest, but varies across trials. ***(c)*** *The MCD model has sequential filtering operations on sensory inputs which lead to a signal that represents the temporal cross-correlation between inputs (MCD_{Corr}). This signal is assumed to accrue Gaussian noise, and a single criterion is applied, such that trials yielding values of (noisy) MCD_{Corr} above this criterion are judged simultaneous. (Note that the x-axis of the 3D inset differs from parts a and b – the relationship between objective SOA and MCD_{Corr}, which is not shown, is roughly inverse-U shaped)* ***(d-f)*** *Solid black lines show resulting simultaneity functions. In each case, the point calculated in parts a-c is highlighted. Other points on the function are obtained in the same way. Dashed black lines show what happens if parameters describing decision criteria are changed to model more conservative behaviour. Dotted grey lines show predictions if criteria are assumed to reflect a hard threshold for the perception of synchrony, so cannot be changed, but participants still attempt to reduce their use of the synchronous response.* ***(g)*** *Replot of data from parts d-f illustrating how a hard-threshold account predicts a linear relationship between proportion judged simultaneous in Baseline and Conservative conditions (regardless of further modelling assumptions) whereas models in which decision criteria change generally predict non-linearity. See main text for further details.*

The first two models come from a family previously labelled “independent-channels” (Sternberg & Knoll, 1973) or “general-threshold” (Ulrich, 1987) models. The core idea is that modality-specific signals (for example a visual flash and an auditory beep) generate neural responses that must propagate through the brain toward a decision centre. As a result, a noisy and delayed version of each signal ultimately arrives at the decision centre. The difference in their subjective arrival times forms the decision variable which must be classified to form a response. For any given experimentally presented asynchrony (objective SOA in Figure 1) it has a distribution whose shape depends on the nature of the latency noise. A “simultaneous” decision is made if the subjective asynchrony falls between two criteria (for example above -100 ms and below 100 ms). One of the central notions behind this family of model (that decision noise reflects latency noise) has recently received support via the recording of simultaneity judgments alongside concurrent electroencephalography (Yarrow et al., 2022). The two models from this family used here, which are outlined next, differ in terms of how they explain asymmetry in the simultaneity function.

Approximation To A Gaussian Latency And Noisy Criteria Equation model of simultaneity

The first model, which we term “AT-A-GLANCE” (Approximation To A Gaussian Latency And Noisy Criteria Equation; Yarrow et al., 2011) assumes that latency noise – trial-by-trial changes in the time taken for the neural responses to propagate through the brain to the decision centre – is Gaussian in shape. On its own, this form of noise generates a symmetric simultaneity function. However, it is further assumed that decision criteria are not held perfectly stable, but rather vary from trial to trial (Ulrich, 1987), introducing a further source of noise that can differ for the two sides of the psychometric function. If criterion noise is greater for one side of the psychometric function (for example when discriminating simultaneous from sound-lags-light stimuli) than for the other (for example discriminating simultaneous from sound-leads-light stimuli) the slope of the function will be flatter on that side.

In order to make it possible to identify the most likely set of model parameters, four sources of conceptual noise (latency noise for each of two stimuli and decision noise at each of two criteria) are combined/reduced into just two noise parameters. These each represent the sum of both sources of latency noise and one of the two sources of criterion noise. Hence this model typically uses a minimum of four parameters per participant/condition, two criterial parameters that determine its position and width and two noise parameters that determine ascending and descending slopes of the psychometric function. Additional parameters may be added for consideration of attention lapses and/or keying errors by the participant. At the time of writing, the effect of changing model parameters can be examined via an interactive Shiny app at https://kielanyarrow.github.io/MyPage/Code.html (see methods for further details of code/data sharing). Previous applications of this model include investigating dissociations between judgments of causality and judgments of simultaneity (Bonnet et al., 2022). It has also helped to account for the phenomenon of temporal recalibration, whereby repeated exposure to a non-synchronous input biases judgments about subsequent stimuli, consistent with participants developing a new impression of what feels synchronous (Yarrow et al., 2013; Yarrow et al., 2015).

Exponential Latency Alone model of simultaneity

Our second model also hails from the independent-channels family (García-Pérez & Alcalá-Quintana, 2012a). We term it "ELA" (Exponential Latency Alone). Rather than assuming Gaussian latency noise, this model assumes that each signal's propagation times through the brain can be better described using an exponential distribution. A judgment is again formed at the hypothetical decision centre by placing bounding criteria on the resulting distribution of subjective differences in arrival times. However, unlike AT-A-GLANCE, these criteria are stable across trials. If each signal gives rise to a different exponential distribution of arrival times (for example the distribution is tighter for auditory than visual signals) this leads to asymmetry in the resulting psychometric function. Leaving aside lapses, this model also uses four parameters per participant: A rate parameter for each exponential distribution, which affect the slopes of the simultaneity function, and two parameters determining its position and width. When each participant completes both simultaneity and temporal order judgment tasks, a simultaneous fit of this model to all tasks at once has been shown to provide a viable account of behaviour (García-Pérez & Alcalá-Quintana, 2012a; García-Pérez & Alcalá-Quintana, 2015). The model has also been used to show how the inclusion of lapse and keying error parameters can allow independent-channels models to deal with findings from ternary tasks (which have before/simultaneous/after response options) that initially appeared to contradict this general architecture (García-Pérez & Alcalá-Quintana, 2012b).

Multisensory Correlation Detector model applied to simultaneity judgments

The final model we implement here has a different background. This MCD (multisensory correlation detector) model (Parise & Ernst, 2016) is broadly analogous to popular accounts of motion detection in vision (Fujisaki & Nishida, 2007). It builds on earlier ideas that perceived simultaneity might be a function of the degree of overlap between the internal responses to two stimuli, which can be thought of as temporally low-pass filtered versions of the input (Burr et al., 2009; Stelmach & Herdman, 1991). In the MCD model, each signal first passes through a modality-specific filter. The output of one modality is then multiplied by an additionally filtered version of the other, and vice versa. Finally, the two resultant signals are multiplied together and then integrated over the interval of time immediately following presentation of the stimuli in order to provide a single quantity (MCD_{Corr}) that represents perceived synchrony. To yield a categorical response, this quantity is compared to a single criterion, above which synchrony is reported. Noise for this judgment accrues from Gaussian variation in either the strength of MCD_{Corr} (which is otherwise deterministic) or the placement of the criterion across trials (these two ideas yield identical predictions so cannot be discriminated).

Leaving aside lapses, this model has five parameters (three filter time constants, a criterion, and a noise term). However, it has traditionally been fitted to simultaneity judgments by fixing the filter time constants based on additional data sets and utilising a two-parameter generalised linear model. Based on our explorations regarding the recoverability of model parameters, we opted to build upon a three-core-parameter (plus lapses) fit. We fixed both the second-stage filter time constant and the visual-filter time constant, but allowed the auditory-filter time constant to vary. Changing the ratio of time constants for the two unisensory filters generates asymmetry in the psychometric function (and also a correlated shift in its central tendency) while the noise term affects slopes, and the criterion term affects width. The model can be explored via our aforementioned Shiny App. Example applications of this model include explaining data from a range of synchrony tasks with stimuli that employ both simple and complex temporal profiles (for example simultaneity judgments, temporal-order judgments, and various judgments about correlated and uncorrelated trains of stimuli). The model has also helped provide viable neural loci for the process of cross-correlating multisensory stimuli (Pesnot Lerousseau et al., 2022) and, with slight modification, helped explain the effect of visual luminance on simultaneity judgments and temporal-order judgments (Horsfall et al., 2021).

Testing whether strategy influences the simultaneity function

Having summarised the candidate models, we can now move on to introduce an experimental manipulation. In previous sections we have alluded to the idea that categorical reports (for example “simultaneous”) might be generated by applying decision criteria to underlying perceptual representations that are continuous. The underlying representation could be an arrival-time difference (as assumed in the AT-A-GLANCE and ELA models) or a cross-correlation of filtered inputs (as per the MCD model). Conscious experience could reflect these continuous quantities, but making binary decisions would require that the underlying representation is categorised using some rule.

However, experience of simultaneity *may* be truly discrete, such that when stimuli are (intrapsychically) close enough in time, or lead to a strong enough simultaneity signal, perception becomes categorically “synchronous” without further nuance (e.g. Venables, 1960). The mind would be like a teacher who, having recorded that a student scoring over 80 receives an A grade, then shreds the test, losing the exact score. To extend the analogy – the cut point for this decision (a score of 80) is not optional, but has been imposed by an exam board. We refer to this kind of mechanism as a hard or structural threshold. Presumably, in the brain it would depend on thresholding mechanisms such as the synapse.

It is straightforward to test whether the criteria applied to the simultaneity-judgment task when participants first walk into the laboratory are hard thresholds of this kind. We can do it, for example, by introducing a condition in which participants are asked to reduce their use of the simultaneous response option (Yarrow, 2018). The models that we have described include parameters which could be allowed to change in such a condition in order to represent a change of decision criteria. This is illustrated by the dashed black lines in Figure 1 parts d-f.

But how would we know that decision criteria had really changed, rather than merely seeming to change as an artefact of fitting an inappropriate model to data? The answer involves predicting what would happen if thresholds obtained at baseline remained hard. With a "be conservative" instruction encouraging a limited number of synchronous responses over the many trials of the experiment, observers would sometimes need to report asynchrony despite perceiving synchrony. The result would be a proportional reduction of the predicted psychometric function (Figure 1 d-f, dotted grey lines). It is straightforward to embed such an account in an observer model, as an alternative parameter that can change in conservative conditions instead of decision criteria.

One concern with such an approach would be that it involves comparing two variants of an observer model, and such models are mere approximations of reality. For this reason, we additionally consider a test of the hard-threshold account that does not depend on any particular observer model. To this end, we can reframe how we visualise the data. Rather than plotting proportion judged synchronous in both baseline and conservative conditions against SOA (as per Figure 1 panels d-f) we can consider proportion judged synchronous in the conservative condition as a function of proportion judged synchronous in the baseline condition (Figure 1g). If the thresholds obtained at baseline are structural, any proportional reduction in judgments of synchrony in the conservative condition (occurring as a result of inferring a need to report asynchrony on a random subset of trials categorically perceived as synchronous) would then translate to predicting a function that is linear on these axes. It would have an intercept of zero and slope equal to the proportional reduction from baseline to conservative conditions.

We now have all the background required to frame our current approach and predictions. In our experiment, participants will initially make simultaneity judgments with limited instruction. This condition evaluates typical/free behaviour when faced with the simultaneity-judgment task. Next, participants will be asked to “be conservative”. For good measure, we will include a final condition in which the instruction is revoked, so as to seek evidence that any changes really were a result of the conservative instruction, rather than, say, practice or fatigue.

First, we will test for an anticipated violation of linearity in the function predicting proportion judged synchronous in the conservative condition from proportion judged synchronous in the baseline condition. Next, for each of the three simultaneity-judgment models we have outlined, we will fit two multilevel model variants to data from all participants and all conditions at once. In the first, parameters relating to decision criteria will be allowed to vary across conditions. In the second, a hard-threshold account will be implemented by instead introducing multiplicative change parameters. We anticipate better fits (when taking into account the number of model parameters) for the former model variants compared to the latter, which would further support the idea that simultaneity judgments are in part strategic. We will also compare goodness of fit across our three types of simultaneity-judgment model (AT-A-GLANCE, ELA, and MCD). Few if any comparisons of this type exist, so we are interested to see which of these models provides a prediction that is closest to our data, and thus receives greatest support.

Method

Participants

This study comprises a secondary analysis of data published previously as a pre-print (Yarrow & Roseboom, 2017). An opportunity sample of twenty observers, all naïve to the experimental purpose, participated in early 2017. Written informed consent was acquired from all participants prior to the experiment, which was approved by the University of Sussex ethics committee. Participants received £5 per hour or course credit as compensation for their time. Demographics were not retained with the dataset, but the sample was recruited from the same predominantly undergraduate student panel, at around the same time, as that reported in Roseboom (2019), which contained 60% females with a mean age around 22 (SD 5) years.

The current work addresses both the originally intended research question (the effect of strategy on simultaneity judgments), but via a more comprehensive analysis, and an additional research question (by comparing different models of the simultaneity judgment). To our knowledge, the most relevant prior observation regarding the effect of strategy came from a single-case study (subsequently described in Yarrow, 2018). This indicated an effect that was large in absolute terms but, with N = 1, could not be normed to a standardised measure of effect size. Hence the sample size was selected (prior to the initiation of data collection) based on prevailing norms for simultaneity-judgment studies with similar designs. Data from one participant were not included in the final analysis (see data analysis, below). For a paired-samples t-test, the remaining N = 19 participants yield a-priori power of 91% to detect a large (Cohen's d = 0.8) effect size (at two-tailed alpha = .05). With regard to our second research question, relating to model comparison, we provide data relevant to evaluating power in Appendix E, where 5/6 simulations (using our sample size) yielded a significant difference between the generative model and each of the non-generative models.

Apparatus and stimuli

Participants sat in a quiet, bright room. Visual stimuli were displayed on either an Iiyama Vision Master Pro 203 or LaCie Electron 22 Blue II monitor, both with a resolution of 1024 x 768 pixels and refresh rate of 100 Hz. The monitor was positioned at a viewing distance of approximately 57 cm. Audio signals were presented binaurally through Sennheiser HDA 280 PRO headphones. Stimulus generation and presentation was controlled through Psychtoolbox 3 (Brainard, 1997) run in MatLab (Mathworks, USA) on a desktop PC. Participants responded using the computer keyboard.

Visual events were luminance-modulated Gaussian blobs (σ = 1.5 degrees of visual angle (dva)) displayed against a grey background (approximately 38 cd/m^2). Peak blob luminance was approximately 76 cd/m^2. A fixation square (white, approximately 76 cd/m^2, subtending 0.25 dva) was presented centrally. The Gaussian blob was centred 3 dva above the fixation square. The visual stimulus was presented for one frame approximating 10 ms in duration. Auditory signals were a 10 ms amplitude pulse of 1500 Hz sine-wave carrier at approximately 55 db SPL.

Design and procedures

The experiment consisted of six sessions. Each took approximately seven minutes to complete. In each session, participants were presented with a sequence of 135 audio-visual presentations. Each presentation consisted of visual and auditory events presented with one of nine pseudo-randomly interleaved stimulus-onset asynchronies (SOAs; $\Delta t \in \{-400\text{ ms}, -200\text{ ms}, -100\text{ ms}, -50\text{ ms}, 0, 50\text{ ms}, 100\text{ ms}, 200\text{ ms}, 400\text{ ms}\}$, where positive values indicate visual stimulus before audio). Each SOA was presented 15 times and preceded by a uniform-random period between 500 ms and 1500 ms. Participants were required to provide an unspeeded response as to whether the auditory and visual events had occurred at the same time (synchronously; up cursor key) or not (asynchronously; down cursor key).

For the first two experimental sessions (270 trials), these were the only instructions given. Before the third and fourth experimental sessions, participants were told: “Be conservative in your responses. Only press the ‘synchrony’ key if you are certain”. No further guidance was given. Following these two sessions, participants completed two further experimental sessions without any limitations on their responses – the same as the first two sessions completed.

Data analysis

Modelling approach and software

We opted to apply Bayesian multilevel models, which we consider the most principled way to treat these data and test our hypotheses. In recent years, multilevel models have seen widespread advocacy and adoption across diverse fields including neuroscience (Aarts et al., 2014) and psychology (Barr et al., 2013). This includes the active promotion of their use to analyse data from psychophysical tasks (e.g. Moscatelli et al., 2012). For standard (sigmoidal) psychometric functions, packages such as the Palamedes toolbox (Prins & Kingdom, 2018) offer multilevel approaches “off the shelf”. However, we are not aware of any such option for those interested in modelling simultaneity judgments. We therefore fit Bayesian multilevel models using the Stan programming language interfaced from R (R Core Team, 2021) via the RStan package (Stan Development Team 2020; 2022). We share our commented code (see Transparency and Openness subsection, below) as a potential template for other researchers interested in developing bespoke multilevel analyses of their own data. Additional R packages including shinystan and LOO were used to diagnose and evaluate models. We fit models using four chains, each exploring the likelihood surface via the default Hamiltonian Monte-Carlo no U-turn sampling (HMC NUTS) algorithm, which retains samples in proportion to the height of the posterior distribution, and thus estimates it. All our reported model fits use 1000 warmup iterations followed by 10,000 post-warmup iterations per chain.

Initial data formatting

Prior to further analysis, we excluded one participant because their adjustment to the instruction to “be conservative” was to significantly *increase* their use of the synchronous response (198/270 vs. 152/270, $\chi 2[1] = 17.98$, $p < .001$), suggesting they had misunderstood the instruction. Data from the remaining 19 participants were summarised as proportion judged simultaneous at each SOA and in each condition. We passed dummy codes for the conservative condition and the post-conservative (rebound) condition to our models, such that the initial uninstructed condition became the baseline for pairwise comparisons.

Assessing group changes across conditions, comparing hypotheses, and considering individual participants

In our model-based analyses, we utilised three classes of multilevel simultaneity-judgment model, each with two variants: A strategic variant which allows one or more parameters that represent participants’ decision criteria to change across experimental conditions, and a hard-threshold variant which instead allows the psychometric function to show proportional reduction. This proportional reduction mimics an attempt to reduce use of the simultaneous response option when all stimuli judged simultaneous give rise to the exact same perceptual experience, as the only option for the participant would then be to reply “asynchronous” at random to some stimuli they perceived as synchronous. Full mathematical details of the models are provided in Appendix A.

These models all incorporate parameters that are conceptually akin to regression coefficients as they quantify the effect of our experimental conditions. They are hence termed β. In assessing whether behaviour changes in the conservative and rebound conditions relative to baseline we are therefore essentially asking whether the group means of the relevant β coefficients differ from either zero or 1.0 – the values that would imply no change from baseline for models of the strategic and hard-threshold accounts, respectively. In a multilevel model, the group mean of individual β coefficients is already estimated as part of the model-fitting process. Hence, in the

Bayesian case, the comparison of this value against zero or 1.0 can be achieved by examining the posterior distribution for the group-level mean (μ_β) coefficients. We provide statements of significance similar to frequentist null-hypothesis testing based on whether the 95% credible interval contains 0 or 1.

We also incorporated *posterior predictive checks* (Lambert, 2018). The posterior predictive distribution of any one of our β coefficients tells us what we can expect for future participants, and in combination with its standard error (which equals its $\mathrm{SD}/\sqrt{\mathrm{N}}$) it can provide an alternative means of evaluating differences from 0 or 1, via a single-sample t-test.[2] We also used a posterior predictive check to evaluate the fit of individual participants by calculating a *Bayesian P value* (Lambert, 2018) representing the proportion of samples for which the likelihood of each participant's actual data was lower than that for a random binomial draw conditioned on model parameters. If the model is correct for an individual, this Bayesian P value should be around .5, with higher values indicating overdispersion and therefore a potentially incomplete or erroneous model. This is conceptually similar to the frequentist approach of comparing deviance of model fit to a chi-square distribution.

Finally, we wished to compare the two model variants (for each class of simultaneity-judgment model) to one another in order to evaluate which of our hypotheses received greater support. We can estimate a model's out-of-sample goodness of fit via leave-one-out cross validation, but this is very time consuming. Hence, we instead used an estimate of leave-one-out cross validation via Pareto smoothed importance sampling (Vehtari et al., 2017), known as PSIS-LOO. This measure is based on the log-likelihood of the model given the data, but utilises the full posterior distribution of parameter values in estimating goodness of fit, and corrects for the number of model parameters in a more nuanced fashion than better-known metrics such as the Akaike and deviance information criteria (AIC, DIC).

[2] At least as long as a consideration of its shape and the sample size N suggests that the sampling distribution of its mean will likely be normal, via the central limit theorem.

PSIS-LOO was estimated and compared between model variants (and indeed between classes of model) using functions from the R package, LOO, and z tests (which are based on the difference between models in units of the standard error of this difference). Although PSIS-LOO can be multiplied by -2 to give an AIC-like value where low is best, we don't bother to apply this transform, so report negative values, where higher is better. PSIS-LOO for the whole model is found by summing log likelihood estimates for each data point. The LOO package provides diagnostics and outputs which together indicate the number and positions of data points for which the PSIS-LOO estimate is potentially inaccurate. We therefore replaced a small number of data points considered "very bad" (Pareto k value > 1.0) via direct leave-one-out cross validation, and also report the number of estimates considered "bad" (Pareto k > 0.7), which we elected not to replace due to the heavy computational demands of doing so.

Transparency and Openness

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. This study's design and its analysis were not pre-registered. All analysis code and data, including shiny apps, is permanently available at https://city.figshare.com/articles/software/Code_and_data_accompanying_The_best_fitting_of_three_contemporary_observer_models_reveals_how_participants_strategy_influences_the_window_of_subjective_synchrony_/20495652.

Results

Non model-based assessment of the hard-threshold account

Figure 2 shows data from the first two conditions of the experiment averaged across participants in two different formats – firstly (panel a) with proportion judged synchronous plotted separately for the baseline and conservative conditions as a function of the time between the visual and auditory stimuli (SOA), and secondly (panel b) with proportion judged synchronous in the conservative condition plotted as a function of proportion judged synchronous in the baseline condition.

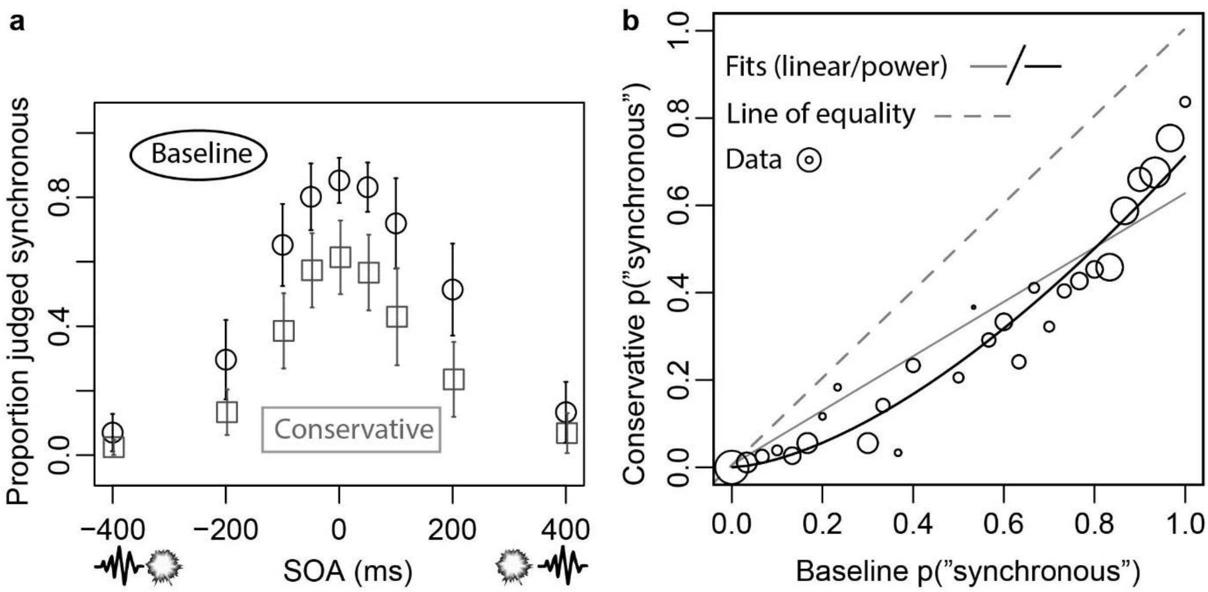

Figure 2. Non SJ-model-based test of the hard-threshold account, focussing on data from the baseline and conservative conditions. ***(a)*** *Error bars show ±2.1 standard errors around the group mean.* ***(b)*** *Size of data points reflects number of participants contributing to each. See main text for further details.*

From Figure 2 panel a, it appears that participants reduced their use of the synchronous response option when asked to "be conservative", but not in a manner that was proportional across stimulus-onset asynchronies. This is confirmed in panel b, where open circles show the group-mean data. It was produced based upon one x/y pair for each SOA and participant (so 19x9 data points in all) from which values with the same baseline proportion judged synchronous were first averaged for each participant, and then across the group. The dashed grey line shows the prediction if there is no reduction in use of the synchronous response. The solid grey line shows the linear prediction for a proportional reduction. This is expected if participants experienced categorical percepts based on an identical hard threshold in the two conditions, but "be conservative" instructions led them to respond "synchronous" on only a random subset of their synchronous percepts.

A linear prediction equates to predicting a power function with an exponent of 1. We therefore sought evidence to reject this null hypothesis by fitting a multilevel model with a zero intercept, but fixed and random effects for both slope and, critically, the exponent of the power function. This specification allows variation in both slope and exponent for each participant.[3] It yielded an estimated group-mean exponent of 1.58 (solid black line in Figure 2b) with a credible interval (1.33-1.85) that did not include 1. This result provides grounds to reject the hard-threshold account. We next moved to more fully characterise our data via three observer models of simultaneity-judgment behaviour, starting with AT-A-GLANCE.

The AT-A-GLANCE model

The AT-A-GLANCE model posits audio and visual signals propagating toward a decision hub, each having Gaussian latency noise. Their subjective difference in arrival times is then categorised using a pair of decision criteria that vary randomly from trial to trial. We fit a multilevel "criterial" variant of the AT-A-GLANCE model to behaviour in all three conditions at once. Multilevel models add a set of group-level parameters to a "heterogeneous" foundation (essentially, a single-level model fitted to each participant). In this case, the heterogeneous foundation specifies a binomial

[3] We used a binomial data model, so very slightly corrected the prediction (to be, at the individual level, $y = 0.00001+0.99999*\text{slope}*x^{exponent}$) to ease likelihood calculations where the model would otherwise predict a be conservative p("synchronous") of zero. This multilevel model assumed Gaussian-distributed group-level parameters (with (improper) uniform hyperpriors for the group's means and standard deviations).

distribution (with 30 trials) for the number of “simultaneous” responses ($S_{X\Delta t}$) from each participant in each condition ($X = \mathrm{B}$, $X = \mathrm{C}$, and $X = \mathrm{R}$, for baseline, conservative, and rebound conditions, respectively) with each objective SOA (Δt):

(1) $S_{X\Delta t} \sim B(30, l + p_{X\Delta t} - l p_{X\Delta t})$,

where l is a free parameter representing (half) the lapse rate with which a participant is distracted and therefore guesses a response and

(2) $p_{X\Delta t} = \Phi\left[\frac{\Delta t - \tau + \beta_{\tau X} D_X - \exp(\beta_{\delta X} D_X)\Delta\delta/2}{\sigma_{\mathrm{L}}}\right] - \Phi\left[\frac{\Delta t - \tau + \beta_{\tau X} D_X + \exp(\beta_{\delta X} D_X)\Delta\delta/2}{\exp(m)\sigma_{\mathrm{L}}}\right]$.

In Equation 2 exp is the exponential function, Φ is the standard normal cumulative distribution function, and D_X is a “dummy” or indicator variable that equals 1 if and only if $X = \mathrm{C}$ or $X = \mathrm{R}$. The remaining 8 symbols (τ, $\Delta\delta$, $\beta_{\tau\mathrm{C}}$, $\beta_{\tau\mathrm{R}}$, $\beta_{\delta\mathrm{C}}$, $\beta_{\delta\mathrm{R}}$, σ_{L}, and m) are all free parameters described below. That makes 9 free parameters for each of 19 participants; a total of 171 for the group as a whole.

In this model, the τ and $\Delta\delta$ parameters capture the midpoint and width (respectively) of each participant’s psychometric function. They provide an alternative (and mathematically equivalent) way of describing the positions of two decision criteria (because $\Delta\delta$ is the distance between these criteria, which are centred on τ). Hence, our hypothesis that decision criteria vary with task instructions can be tested by allowing these two parameters to vary across conditions. For this purpose, four parameters, $\beta_{\tau\mathrm{C}}$, $\beta_{\tau\mathrm{R}}$, $\beta_{\delta\mathrm{C}}$ and $\beta_{\delta\mathrm{R}}$, represent changes between conditions (compared to baseline) with the first subscript representing the parameter being adjusted and the second representing the Conservative and Rebound conditions. The σ_{L} and m parameters describe noise affecting the left flank of the psychometric function, and the noisiness of the right flank relative to the left flank (m of 0 indicating an identical magnitude of noise), respectively. Like the lapse-rate parameter l, these final two parameters were assumed constant across experimental conditions.

For our second "hard-threshold" AT-A-GLANCE model variant, the four parameters permitting changes across conditions were replaced with just two (β_{C} and β_{R}), each describing a proportional reduction in the number of trials judged synchronous for a given condition.

For both variants, our multilevel models additionally estimated random variation across the group via group-level distributions from which the individual-level parameters were drawn. This required a further 17 (or 13) parameters (for criterial and hard-threshold variants, respectively). For example, we estimated, for the Gaussian group-level distribution of individual τ parameters, a group mean (μ_τ) and standard deviation (σ_τ). Similarly, for the group-level distribution of *changes* in τ from the baseline to the conservative condition, we estimated a further group mean ($\mu_{\tau\mathrm{C}}$) and standard deviation ($\sigma_{\tau\mathrm{C}}$). Full details are provided in Appendix A (with group-level distributions visualised in Appendix C).

We carried out a number of checks to verify that our modelling procedures were sensible. These indicated that AT-A-GLANCE's posterior likelihood surface was recovered adequately (Appendix B). Furthermore, our design choices for priors and hyperpriors did not appear to exert untoward influence on our conclusions (Appendix C). Finally, we were able to successfully recover parameters for simulated data (Appendix D).

Figure 3 presents the fit of the criterial AT-A-GLANCE multilevel model for all participants in all three conditions. Assessed by eye, the model appears to be capturing the data well, including trends across conditions in response to changes of instruction.

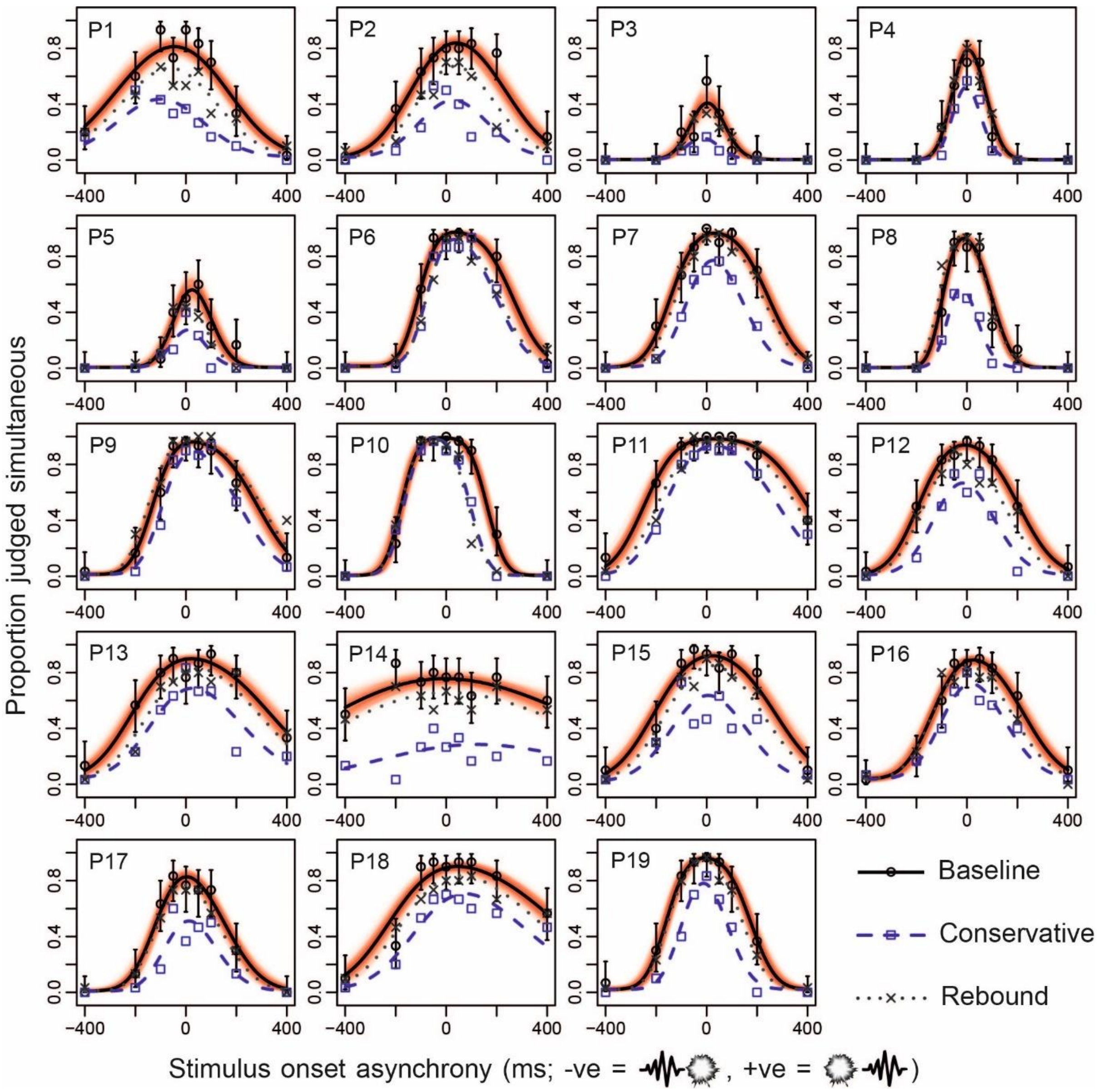

Figure 3. Predictions (based on means of posterior parameter distributions) for the AT-A-GLANCE (criterial-variant) model, alongside data, for all 19 participants in all three conditions of the experiment (Baseline, Conservative, and Rebound). Exclusively in the Baseline condition, red background shading has been added to represent 1000 samples from the full posterior (each plotted with high transparency) in order to illustrate uncertainty in the model prediction, and error bars (which represent 95% binomial confidence intervals) have been added to illustrate uncertainty in the data.

Table 1 summarises the two variants of each of our three models. Focussing on the first two rows, we can see that goodness of model fit, quantified by the PSIS-LOO metric, is better (i.e. PSIS-LOO is higher) for the multilevel variant of AT-A-GLANCE that allows criteria to change across the three conditions (illustrated in Figure 2) than for the alternative hard-threshold variant, which assumes that the categorical boundaries demarcating judgments of synchrony from asynchrony cannot be changed at will.

Table 1. Summary of models.

Model	**< Parameters >**		**< Goodness of fit >**		
	Total	**Group-level changes from baseline condition captured using:**	**PSIS-LOO**	**% Pareto k 0.7-1**	**N dispersion P > .95**
AT-A-GLANCE criterial	**188**	$\mu_{\tau C}, \sigma_{\tau C}, \mu_{\tau R}, \sigma_{\tau R}$ $\mu_{\delta C}, \sigma_{\delta C}, \mu_{\delta R}, \sigma_{\delta R}$	**-1071.2**	**2.5**	**1**
AT-A-GLANCE hard threshold	**146**	$\varphi_C, \lambda_C, \varphi_R, \lambda_R$	**-1129.1**	**1.2**	**6**
ELA criterial	**188**	$\mu_{\tau C}, \sigma_{\tau C}, \mu_{\tau R}, \sigma_{\tau R}$ $\mu_{\delta C}, \sigma_{\delta C}, \mu_{\delta R}, \sigma_{\delta R}$	**-1115.9**	**4.5**	**3**
ELA hard threshold	**146**	$\varphi_C, \lambda_C, \varphi_R, \lambda_R$	**-1155.5**	**1.6**	**5**
MCD criterial	**125**	$\mu_{CC}, \sigma_{CC}, \mu_{CR}, \sigma_{CR}$	**-1156.7**	**1.6**	**6**
MCD hard threshold	**125**	$\varphi_C, \lambda_C, \varphi_R, \lambda_R$	**-1152.5**	**1.8**	**4**

PSIS-LOO is similar to better-known metrics such as AIC in that it approximates a model's out-of-sample predictive capability (specifically the log-likelihood that would be obtained via leave-one-out cross validation). Like all such approximations, it depends on assumptions. For PSIS-LOO (unlike many alternatives) assumptions are conveniently tested alongside its calculation. They are violated when data points yield a high value of a metric called Pareto k. We therefore directly determined leave-one-out log likelihood for data points with very worrisome values of Pareto k (above 1), and also report the percentage of somewhat worrisome data points (Pareto k 0.7-1) as a guide to possible error in the PSIS-LOO approximation. Table 1 indicates that any such error was small.[4]

We can therefore reasonably compare PSIS-LOO values between the two model variants that formalise different theories regarding how participants respond to instructions across our three experimental conditions. The difference in PSIS-LOO of 57.9, with a standard error of 19.9, implies that the criterial AT-A-GLANCE model fits the data considerably better (frequentist two-tailed z test, $z = 2.91$, $p = .004$). This gives us confidence to assert the following: If AT-A-GLANCE is a reasonable approximation of the processes underlying synchrony judgments, participants generally seem able to make adjustments to a pair of internal criteria for simultaneity in order to moderate their use of the synchronous response.

Some evidence that criterial AT-A-GLANCE is in fact a plausible account of these data (in an absolute sense) comes from considering our Bayesian P values. These quantify, for each participant, the degree of overdispersion (meaning residual errors greater than implied by the format of the data, so here, higher than a binomial distribution would be expected to yield).[5] As indicated in Table 1, for the criterial model, only one out of 19 participants had a Bayesian P value above .95, which is around the chance expectation if the model is correct. However, for the hard-threshold model, 6 participants showed overdispersion of this magnitude.

In Figure 4, we plot results from a subset of four participants – those showing the lowest and highest overdispersion, so effectively the best and worst fits, for each of the two different variants of the AT-A-GLANCE multilevel model. The hard-threshold model cannot capture a common pattern in

[4] The AT-A-GLANCE criterial model's estimate of leave-one-out log likelihood may be slightly off (with 2.5% of data points showing Pareto ks of 0.7-1), but when we directly determined leave-one-out log likelihood for values of Pareto k above 1, the maximum error we observed (compared to the PSIS-LOO approximation) across all such data points and all of our models was only around 18%. Data points with Pareto k values of 0.7-1 should, if anything, be better estimated than this, suggesting a misestimation of less than 18% occurring for 2.5% of the overall estimate, implying a fairly small error. For the AT-A-GLANCE hard-threshold model, the error should be even lower.

[5] Technically, our Bayesian P values are the proportion of posterior samples for which the data are more dispersed than a random draw based on the model.

which the psychometric function contracts inwards from one or both sides (participants 10, 19, and several others not shown in Figure 4). The criterial model struggles only with a rarely observed pattern in which the psychometric function shrinks downwards (participant 15).

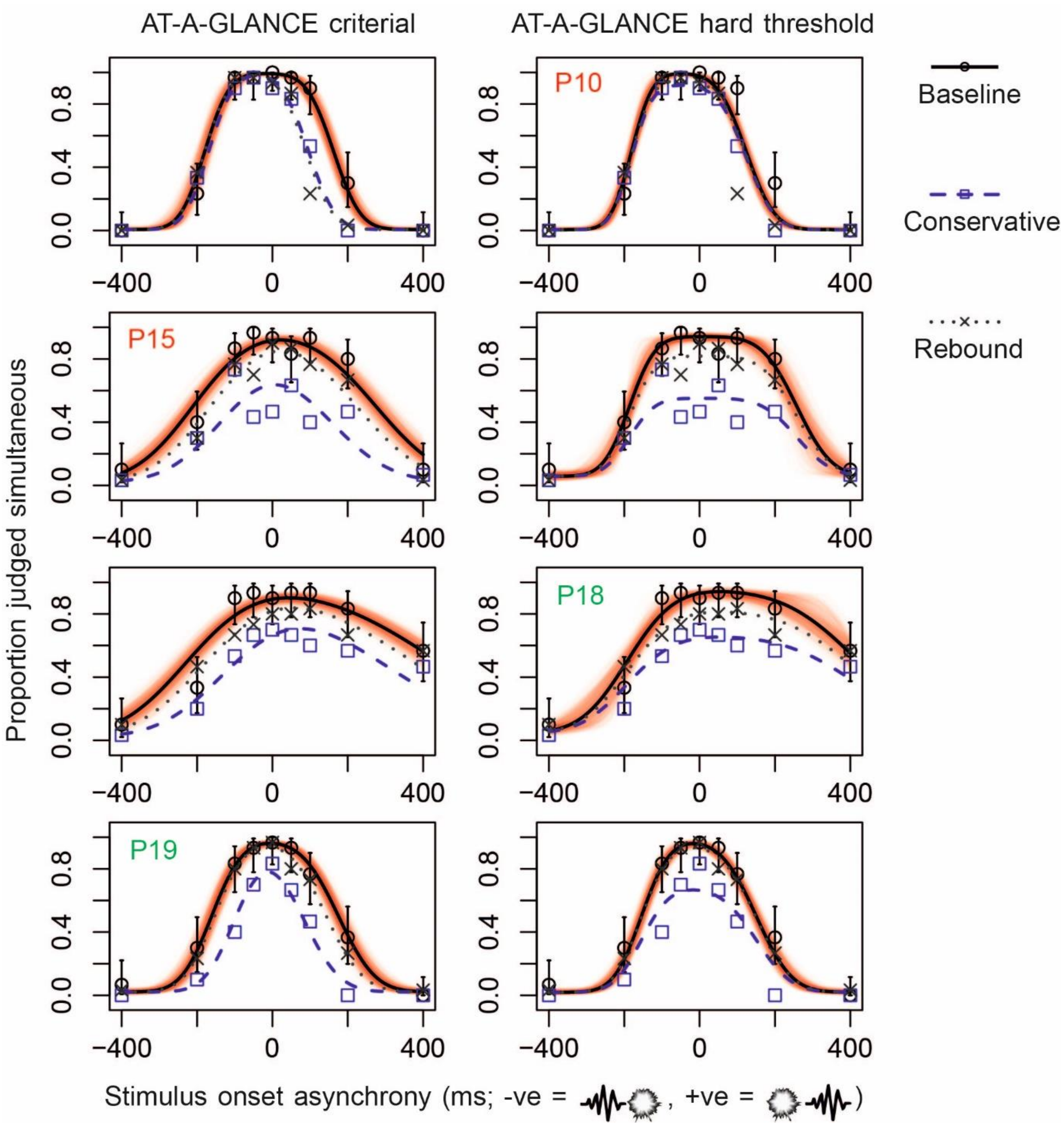

Figure 4. Predictions (based on means of posterior parameter distributions) for both variants of the AT-A-GLANCE model, alongside data, for four illustrative participants in all three conditions. Green text denotes the best-fitting participant for a given model, while red text denotes the worst-fitting participant. Exclusively in the Baseline condition, red background shading has been added to

represent 1000 samples from the full posterior (each plotted with high transparency) in order to illustrate uncertainty in the model prediction, and error bars (which represent 95% binomial confidence intervals) have been added to illustrate uncertainty in the data.

We can also consider exactly how the parameters of the significantly more successful criterial variant of the AT-A-GLANCE model have changed across the three experimental conditions. Two parameters were allowed to change. The first, τ, describes the point midway between decision criteria, and is comparable with the commonly reported "point of subjective simultaneity". In the baseline condition, the mean of its group-level distribution (μ_{τ}) was 32 ms (95% credible interval 10 to 55). This implies a group-average bias to report simultaneity more when sound lags light than vice versa (individual values for all participants can be seen in Appendix C Figure C1a). However, this bias was reduced in the conservative condition (relative to baseline). The mean of the distribution describing *changes* in psychometric function central tendency ($\mu_{\tau C}$) was -23 ms (95% credible interval -35 to -12). This implies a statistically compelling leftward shift of the psychometric function, and highlights how estimates of the point of subjective simultaneity can be affected by participant response strategy. Importantly, we also observed that the mean of the distribution describing changes in psychometric function *width* ($\mu_{\delta C}$) was -0.65 (credible interval -0.80 to -0.50; see also Appendix C Figure C1f). This implies a horizontal contraction of participants' psychometric functions from baseline to conservative conditions which was statistically compelling. For the $\mu_{\delta C}$ coefficient, exponentiation provides more meaningful units: The distance between decision criteria has changed (shrunk) by an average factor of 0.52. This means that participants are making simultaneous responses for a reduced range of audio-visual timings.[6]

Changes in position and width can also be re-expressed in terms of the individual positions of each of two decision criteria, which determine which subjective SOAs will be categorised as

[6] Posterior predictive checks provide near-identical estimates for the mean shift and contraction, and also offer a route to a frequentist test of statistical significance (one-sample t-tests vs. 0; t = 6.72 and t = 8.55 respectively, df = 18, both p < .001).For these and the equivalent t-tests reported subsequently, effect sizes can be easily determined if required as Cohen's *d* = t/√19.

simultaneous. Both have moved inwards in the conservative condition, but this change is less pronounced for the low criterion. It showed an average shift of +70 ms, but with a credible interval from -11 ms to 326 ms that hence includes zero. The high criterion showed an average shift of -115 ms (credible interval -373 ms to -33 ms). Regardless of how the criteria have been parameterized, their shifts suggest that participants appropriately adjusted their decision-making strategies in accordance with the instructions to be more conservative. More specifically, participants made more of an adjustment regarding how light-leading stimuli should be classified compared to how sound-leading stimuli should be classified.

In the rebound condition, relative to baseline, a less pronounced version of the same pattern emerged. The psychometric function shifts left ($\mu_{\tau R}$ = -9 ms, credible interval -20 to 2 ms) and contracts ($\mu_{\delta R}$ = -0.17, credible interval -0.09 to -0.26) by a factor of 0.84.[7] This is equivalent to mean changes to the low and high criteria of 23 ms (credible interval -29 to 98 ms) and -43 ms (credible interval -116 to 20 ms) respectively. As these changes are relative to baseline, this suggests that participants did not completely revert back to their original lax decision criteria.

The ELA and MCD models

In addition to the above-described results for the AT-A-GLANCE model, we tested two further models of the synchrony judgment: ELA, which is similar to AT-A-GLANCE but assumes exponential latency noise and stable decision criteria, and MCD, which infers simultaneity from overlap in neural responses, rather than arrival times at a neurocognitive hub. Mathematical details appear in Appendix A. Returning to Table 1, it is apparent that the AT-A-GLANCE criterial-model variant shows substantially better goodness-of-fit metrics compared to all other models. A statistical comparison suggests that these differences are meaningful. We focus on the generally better-performing criterial variants of each class of model. A difference in PSIS-LOO of 44.7 (with a standard

[7] Posterior predictive tests yielded one-sample t = 2.25, df = 18, p = .037, and t = 4.76, p < .001, for $\mu_{\tau R}$ and $\mu_{\delta R}$ respectively.

error of 14.6) implies that the criterial AT-A-GLANCE model fits the data considerably better than the criterial ELA model (frequentist two-tailed z test, $z = 3.06$, $p = .002$). The difference was even more striking for criterial AT-A-GLANCE versus criterial MCD (PSIS-LOO difference = 85.5, SE 20.5, $z = 4.17$, $p < .001$). Criterial MCD also performing somewhat badly relative to criterial ELA (PSIS-LOO difference = 40.8, SE 21.9, $z = 1.86$, $p = .062$).[8] As a methodological check (and test of model mimicry), we investigated, via simulation, the extent to which the PSIS-LOO metric would have favoured any of our three models in the case where that model was the true data-generating model (Appendix E). AT-A-GLANCE seemed better able to mimic ELA than vice versa, but yielded significantly better PSIS-LOO only when it was the true model, and was on average worse when it was not. The MCD model was beaten convincingly by both AT-A-GLANCE and ELA when they were generative and it was not, but also significantly outperformed them when it was the generative model. These findings imply that the correction for model complexity built into PSIS-LOO worked as intended in the current context.

Because both ELA and MCD provided significantly less compelling descriptions of the data relative to AT-A-GLANCE, we will spend less time describing their detailed results. However, Figure 5 provides some insights into why these models performed less well. The figure plots fits from all three classes of model (specifically their criterial variants) for the subset of participants for whom any model particularly struggled (those with overdispersion Bayesian P values >.95, cf. Table 1).

[8] Given that the MCD model's hard-threshold variant had a higher PSIS-LOO than its criterial variant, it may be fairer to compare against this value. Here, the difference compared to criterial variants of AT-A-GLANCE and ELA was significant ($p = .001$) and non-significant ($p = .204$) respectively.

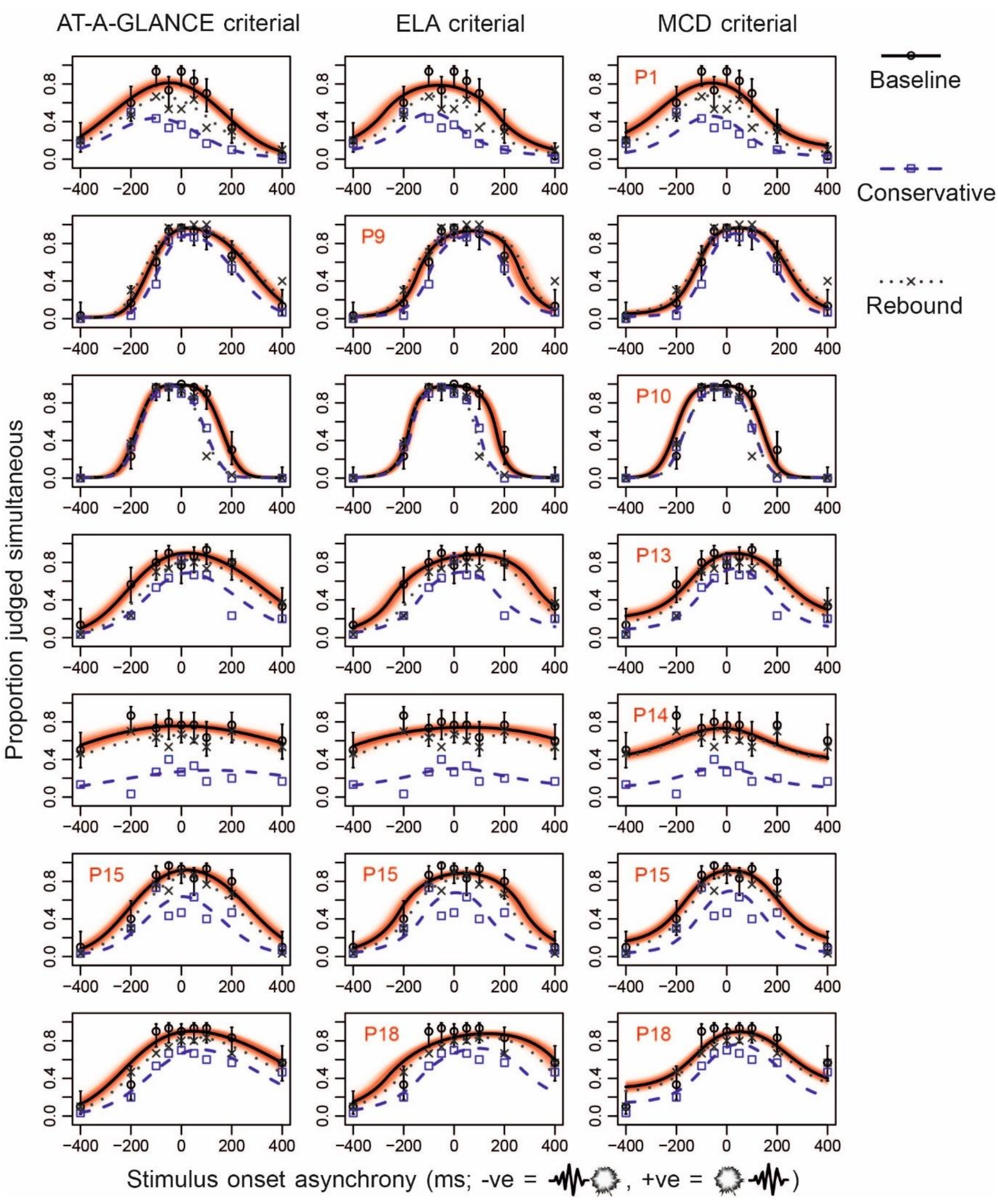

Figure 5. Predictions (based on means of posterior parameter distributions) for the criterial variant of the AT-A-GLANCE, ELA and MCD models, alongside data, for seven illustrative participants in all three conditions. Red text denotes that data were overdispersed (Bayesian P > .95) for that participant/model. Exclusively in the Baseline condition, red background shading has been added to represent 1000 samples from the full posterior (each plotted with high transparency) in order to

illustrate uncertainty in the model prediction, and error bars (which represent 95% binomial confidence intervals) have been added to illustrate uncertainty in the data.

Figure 5 illustrates that, in general, the MCD model had a difficult time accounting for those participants whose conservative adjustment was more notable for sound-lagging than for sound-leading stimuli (for example participant 10). Problems with ELA are trickier to characterise, but seem to reflect constraints on the exact shape of the psychometric function, particularly relating to some participants' poor performance at sound-lags-light SOAs (for example participant 18).

For the ELA class of model, like AT-A-GLANCE, the criterial variant showed better goodness of fit compared to the hard-threshold variant. However, the difference in PSIS-LOO (39.6, with a standard error of 26.9) was not statistically compelling ($z = 1.47$, frequentist two-tailed $p = .141$). For the MCD model, the trend actually reversed (with the hard-threshold variant outperforming the criterial variant) although the magnitude of the difference was very small in relation to estimation error (PSIS-LOO difference 4.2, SE 25.3, $z = 0.17$, $p = .868$). However, because AT-A-GLANCE provided a significantly better overall account of the data, and agreed with our non-model-based test (see section *Non model-based assessment of the hard-threshold account*), we give priority to that model when interpreting differences between model variants in relation to our experimental hypothesis.

Discussion

In this paper, we have presented data from an experiment requiring judgments about the simultaneity of audio-visual pairs. Participants made these judgments under conditions that either let them freely decide how to behave, or encouraged them to be conservative in their use of the simultaneous response option. Data were then fitted with two variants of each of three multilevel observer models of simultaneity judgments. The two variants of each model represented different hypotheses about how participants would attempt to address the experimental instruction. If flexible criteria exist and determine which subjective stimulus patterns are classified as simultaneous, participants would be expected to adjust those criteria when asked to be conservative. If no such criteria were being applied in the first place, consistent with truly binary perceptual experiences arising from some hard neurocognitive thresholding mechanism, participants would have two choices. They might either fail to adjust their behaviour at all, or sometimes respond "non-simultaneous" even to perceptually compelling experiences of simultaneity in order to meet experimental demands.

Our first observation is that of the three classes of model that we tested, AT-A-GLANCE (Approximation to a Gaussian Latency Independent Noisy Criteria Equation; Yarrow et al., 2011), a variant of the general-threshold family of models (Ulrich, 1987) provided the best account of the data. This is, to our knowledge, the first time a direct comparison between two or more of these models has been attempted. Given AT-A-GLANCE's success (in both relative and absolute terms) we prioritised this model for the evaluation of our experimental hypothesis regarding the existence of decision criteria. Of AT-A-GLANCE's two variants, the criterial variant, corresponding with the hypothesis that participants were applying flexible internal decision criteria in order to categorise stimuli as simultaneous or not, significantly outperformed the hard-threshold variant. This was in accord with our non-model-based test, which also provided grounds for rejecting the hard-threshold account. Differences between the baseline condition and the "be conservative" condition (and, to a lesser extent, a subsequent rebound condition) were well accounted for by a shrinking-inwards of two decision criteria applied to the subjective difference in arrival times between auditory and visual signals. The movement of the high criterion (that distinguishes simultaneous from sound-lags-light stimuli) was more pronounced than that of the low criterion (distinguishing simultaneous from sound-leads-light stimuli).

AT-A-GLANCE performed better than ELA, but the wider family of models bears further examination

The most successful of our models, AT-A-GLANCE, has much in common with the second most successful, ELA (García-Pérez & Alcalá-Quintana, 2012a). Both posit signals propagating through the brain toward a decision centre and accumulating latency noise in the process, an idea that has received recent support based on an analysis of simultaneity judgments alongside recordings of EEG (Yarrow et al., 2022). Furthermore, both posit that judgments of simultaneity arise when the subjective difference in arrival times at this decision centre falls within a limited window. The models differ in terms of the forms of latency noise that are envisaged, and whether simultaneity criteria are viewed as being constant or variable from trial to trial.[9]

AT-A-GLANCE's particular combination (Gaussian latency noise and variable criteria) was more successful than ELA's (exponential latency noise with fixed criteria) as a description of the shape of psychometric functions implied by the current data. However, the decision to use exponential latency noise in ELA appears to have been largely a matter of mathematical convenience

[9] It is perhaps worth noting at this point that while we have talked rather loosely in terms of decision criteria for both AT-A-GLANCE and ELA, on our reading, García-Pérez and Alcalá-Quintana have a philosophical preference for the existence of a true hard threshold (which enforces guesses for tasks such as the TOJ). However, nothing about the mathematics of their SJ model imposes this interpretation. They have often allowed their parameter δ (which represents half the distance between decision bounds and appears as $\Delta\delta/2$ in our notation) to vary in joint fits (for example allowing it to differ between TOJ and SJ tasks). This suggests that they may consider at least some judgments of simultaneity to have occurred when a strict (structural) hard threshold beneath which perception becomes categorical has not yet been reached.

(and gives rise to both a computationally efficient model prediction and a posterior likelihood surface that is highly amenable to search and characterisation). Meanwhile, AT-A-GLANCE's use of Gaussian noise must be strictly incorrect to the extent that it permits propagation times to be negative. Something in between the two (for example some shifted gamma distribution aside from the exponential), probably with additional criterion noise, therefore holds conceptual appeal. It would be plausible when considering the nature of neuronal transmission, and offer the possibility of separately characterising the noise associated with each stimulus. However, there are practical issues to consider that make this avenue of research challenging. Model parameters would likely become more degenerate (meaning it would be more difficult to recover a unique value for each). There would also be increasingly subtle differences between the psychometric functions that different blended models would predict.

AT-A-GLANCE performed better than MCD because of core MCD features that may not be amenable to a quick fix

The multisensory correlation detector (MCD) model (Parise & Ernst, 2016) is a highly attractive one. It offers both a lower level of abstraction relative to both AT-A-GLANCE and ELA, and the promise of immediate application to a wider range of experimental tasks, such as those involving complex trains of stimuli. However, MCD was markedly less successful in describing our data set. This might in part be because it does not offer independent mechanisms to affect the central tendency of the simultaneity function and the relative slopes of its two flanks. However, the more fundamental problem seems to have been that under the multisensory correlation detector model the derived decision variable (MCD_{Corr}) effectively throws away information about the sign of the SOA. Hence any change in the (single) decision criterion that is applied to this signal has similar effects at both sides of the simultaneity function. In contrast to this, some participants seem to selectively adjust decisions more for sound-lagging compared to sound-leading stimuli.

It is difficult to envisage a simple change that might resolve this problem, because it arises from a core feature of the MCD architecture. Hence, at this point we conclude that an MCD_{Corr}-like signal cannot be the only source of information determining how participants judge simultaneity in the simultaneity-judgment task (although it might contribute). In saying this we do however acknowledge that there could be systemic differences between how timing decisions are made between different individuals or groups. A focus on group-level summary measures comparing distinct models might obscure any such differences.

Interpretations based on simultaneity judgments should bear in mind the task's criterion-dependent nature

Broadly, there are two mechanisms which might be envisaged as a limit on an observer's precision (or on their sensitivity or acuity, which are synonymous terms). The first is internal noise. The second is an inflexible (hard) thresholding mechanism which irretrievably reduces a continuous representation regarding a perceptual dimension (for example the timing between two events) to a categorical one. A key finding from our experiment is that both a non-model-based test and the best supported model (AT-A-GLANCE) provide converging evidence regarding whether a hard threshold should be inferred from SJ data. Both favour the alternative idea that judgments of simultaneity are formed by classifying a continuous underlying signal according to decisional criteria. The fact that these decisional criteria reverted only partially in the rebound condition suggests that, for many participants, at least three criterial settings were attainable. It might also imply that the settings adopted initially had no special/default status.

Such flexibility implies that the width of the simultaneity function tells us mostly about how conservative or liberal participants are in the application of their decisional criteria regarding the category "simultaneous". This account is consonant with a number of findings. For example, a wider simultaneity function is found when judging synchrony between a sound and a bouncing visual display compared to a streaming visual display (Vroomen & Keetels, 2020). Simultaneity function width is also greater for pairs of stimuli previously encountered as co-occurring compared to pairs that are novel (Habets et al., 2017). Both the percept of bouncing/causality, and semantic or probabilistic knowledge about co-occurrence, likely encourage the use of more liberal criteria for judging simultaneity (see also Roseboom et al., 2009, for increased conservatism caused by temporal clutter).

It is worth clarifying that our arguments here against a hard-threshold account relate specifically to the determinants of *typical* simultaneity-judgment behaviour. They do not rule against the existence of such a structural threshold within (or above, in the case of the multisensory correlation detector model) the criterial range that is naturally obtained. The current methodology might be extended to address this kind of question, or at least to place a limit on the magnitude of any structural threshold, by forcing ever-more conservative behaviour through stricter rationing of the simultaneous response option. Ideally, this would be done with highly motivated participants and closely spaced SOAs. Such an approach would complement previous attempts to test hard-threshold accounts for relative time. For example, Baron (1971) offered a first and second guess about which of two synchronous and one preceding stimulus came first, and assessed the degree to which second guesses (following an initial failure) yielded above-chance performance. That approach, which focussed specifically on triads of intramodal (visual) stimuli, ruled out certain kinds of hard-threshold account (Allan, 1975b). These include accounts in which noise in performance comes relative to a background sampling process (for example the moving moment model of Stroud, 1956). However, it also provided evidence against independent-channels models without any thresholds. With the addition of appropriate model comparison, it might be used to formally assess remaining alternatives, such as models with hard (“low”) thresholds accompanied by sensory noise (Swets et al., 1961). Our results here indicate that if, in the audio-visual case, a hard threshold does exist alongside sensory noise, the request to simply judge simultaneity (without further constraint) does not lead participants to judge synchrony only when that threshold is breached. Hence this combination of instruction and task does not reveal what that threshold might be.

These findings regarding the important role that decision criteria play stand in contrast to the widespread interpretation of simultaneity-function width as an unambiguous measure of the *precision* of multisensory integration (e.g. Chen et al., 2017; Foucher et al., 2007; Habets et al., 2017; Hillock et al., 2011; Lee & Noppeney, 2011; Marsicano et al., 2022; Navarra & Fernández-Prieto, 2020; Noel et al., 2017; Scarpina et al., 2016; Stevenson et al., 2014; Zampini et al., 2005). We have already indicated how our results show that simultaneity-function width in uninstructed baseline conditions is not a measure of a hard sensory threshold, if indeed one exists. That leaves the question of whether it is a measure of internal noise. It is plausible, and even predicted by some accounts of what an optimal observer is trying to do, that there might be a correlation between the spacing of decision criteria and the noise underlying perception. Sensitivity should often inform strategy, potentially linking these conceptually distinct measures (Magnotti et al., 2013). However, researchers should be mindful that any difference between the widths of simultaneity functions would then only be indirectly driven by differences in, for example, the consistency of arrival times at a central comparator. We note that the naïve expectation that wider windows of perceived simultaneity should predict less or worse multisensory integration has received somewhat mixed support (for example Stevenson et al., 2018). Viewing the width of the simultaneity function from our model-based perspective might help explain why.

The fact that perceptual precision and simultaneity-function width can dissociate leads us to argue that there should be wider discussion of this issue. Several groups have demonstrated that the widths of simultaneity functions differ between clinical or special-interest groups and controls (for example those experiencing early visual deprivation: Chen et al., 2017; schizophrenics: Foucher et al., 2007; musicians: Lee & Noppeney, 2011). These remain interesting observations, regardless of *why* they differ. However, we believe researchers should point out that these changes do not necessarily reflect perceptual limitations. Moreover, given that there are easily derived model parameters that have a better claim to represent internal noise in multisensory perception (for example those affecting the slope of the simultaneity function, such as σ parameters for AT-A-GLANCE and MCD, and λ parameters for ELA) we suggest that these measures should more often take the limelight.

If the key interest is not noise in multisensory timing, but instead the range of times across which multisensory signals are integrated/bound, the best approach might be to use a task that measures the researcher's definition of integration/binding, rather than the participant's definition of simultaneity. For example, consider the redundant-signals effect. This is a reaction-time advantage obtained over and above a statistical facilitation when responding to audio-visual pairs rather than their individual components (e.g. Colonius & Diederich, 2004; Diederich & Colonius, 2015; Hershenson, 1962; Miller, 1982; Raab, 1962; Schwarz, 2006). It is measurable when the audio-visual pair is near synchronous. The redundant-signals effect implies multisensory integration has occurred: The two signals have interacted in a way that modifies behaviour relative to the sum of their individual effects. Furthermore, the timing between component signals is an important determinant. It would, in our opinion, be a reasonable task with which to quantify the dependency of multisensory integration upon the timing between signals. By contrast, at least at face value, the range of audio-visual timing relationships over which I declare two signals to be simultaneous has little claim to measure the range of values at which my brain integrates/binds them in order to generate a multisensory advantage. We would argue that the near-ubiquitous (but extremely leading) term "temporal binding window" should be replaced with something more neutral, like "window of subjective simultaneity" when summarising the results of simultaneity-judgment studies.

Bayesian multilevel modelling is a complex but powerful approach to analysing simultaneity judgment experiments

We could have fitted our models using the common two-step approach of first fitting a model to each individual, and then assessing group differences using a procedure such as the *t*-test. Multilevel models have advantages over such a two-stage analysis. Perhaps most importantly, by fitting all participants at once, multilevel models can generate "shrinkage", whereby well-estimated participants help constrain parameter estimates for less well-estimated participants (Lambert, 2018). The result can be more powerful, robust and reliable estimation that generally performs better in out-of-sample prediction (Aarts et al., 2014; Lambert, 2018; Moscatelli et al., 2012). Shrinkage may also have practical value in a field where it is common to reject participants on the basis that their data are inadequate to generate reliable parameter estimates (and in which pre-registration of exclusion criteria is not yet the norm). If there are ways to reduce the number of participants who have to be excluded, we should probably adopt them.

Bayesian models additionally encourage the explicit specification of sensible priors, or rather hyper-priors in the case of multilevel models. When used judiciously, these should further enhance the reliability of recovered parameters. They also make use of the full distribution of plausible parameter values from the posterior when assessing the goodness of a model's fit, rather than relying exclusively on the mode of the posterior, as per maximum likelihood estimation. Compared to popular metrics like the Akaike information criterion (AIC), Bayesian metrics (for example estimation of leave-one-out cross validation via Pareto smoothed importance sampling; Vehtari et al., 2017) are likely to provide a better estimate of a model's out-of-sample predictive accuracy, and thus a fairer means of comparing models with different architectures (Lambert, 2018). Here, we have demonstrated how such Bayesian multilevel modelling can be used to evaluate whether model parameters change across conditions, and to test more complex hypotheses via the instantiation of these hypotheses as competing models.

We hope that the code accompanying this paper, in concert with Appendix A, can act as a template for other researchers interested in using similar approaches. Although we have focussed on the popular simultaneity-judgment task, there is a range of tasks that generate non-sigmoidal psychometric functions that might benefit from bespoke multilevel modelling along these lines. In the realm of time perception, these include judgments about which of two intervals contained a more synchronous signal (Yarrow et al., 2016) or whether the duration of a test stimulus matched that of a pre-learnt standard, often referred to as temporal generalization (Bausenhart et al., 2018; García-Pérez, 2014). There are also analogous tasks in other fields (e.g. García-Pérez & Peli, 2014; Morgan et al., 2013). Nonetheless, we must acknowledge that because of the need for bespoke coding, the time investment for this type of analysis exceeds that associated with the application of simpler tests (such as t-tests) as a second-stage inferential step. For example, we have only illustrated a test of whether/how parameters change across a single experimental factor, via dummy coding. Implementing factorial designs would require technical knowledge regarding how to implement the equivalent of ANOVA models within a multilevel model framework, for example the proper use of effects coding. However, we doubt this is beyond the abilities of the average quantitively minded researcher.

We are additionally mindful that the benefits of shrinkage that accrue from the multilevel approach are premised on the correctness of modelling assumptions regarding group-level distributions. For example, in the AT-A-GLANCE and ELA models we assumed a normal distribution for the group when modelling the τ parameter. This describes the central tendency of individual simultaneity-judgment functions, so is the parameter most conceptually akin to the commonly reported "point of subjective simultaneity". But what if the population actually consists of a number of distinct sub-groups, perhaps reflecting very different task strategies or neurological types? Then, the implied uniformity, in terms of the computational processes underlying timing decisions, would be incorrect, and shrinkage toward the group mean could be inappropriate. This would be most pernicious if differences that lead to poor parameter estimation (and thus maximise the reliance on group-level priors) are more likely for members of distinct minority groups (to whom those priors may not apply). In theory, one might address this with something like a mixture distribution for the prior, but this would be challenging in practice. However, if groups are a priori identifiable (for example via a diagnosis), it would be straightforward to implement a between-participants design factor via discrete group-level distributions.

Further caveats, limitations, and constraints on generality

There are several reasons to be cautious regarding our conclusions here, which are derived from work with a necessarily limited scope. Firstly, our study lacked a fundamental feature of well-designed repeated-measures experiments – the counterbalancing of the order of experimental conditions to remove practice and fatigue effects. This was justified by our desire to capture instinctive behaviour in the simultaneity-judgment task before meddling with people's strategies, but it implies that differences between conditions might be contaminated by learning effects. We acknowledge this problem, but note that the inclusion of the rebound condition provides some reassurance that the main driver of differences between conditions was the instruction we provided.

Secondly, with the exception of our non-model-based test of the hard-threshold account, our conclusions follow from the exact choices we made when implementing simultaneity-judgment models, and strictly cannot be generalised beyond that context. For example, we used a single lapse-rate parameter l, but might reasonably have used two such parameters to capture a bias towards one or other response when guessing, as has been implemented by the authors of ELA (García-Pérez & Alcalá-Quintana, 2012b). We gave all three models identical flexibility in this regard, but it is possible that their relative statuses would have changed had we made different choices. The same follows for other decisions, including our choice of hyperpriors (but see Appendix C) and the parameters that were allowed to change across conditions. Allowing only criteria to vary was largely dictated by the logic of the experiment, but a case could be made for also allowing changes in precision due to learning (although note that we did not provide any feedback). In fact, one consequence of bespoke multilevel modelling is that it discourages the testing of a large number of such variant ideas, because each one must be somewhat laboriously coded. Researchers will probably have differing opinions about whether this is a good or a bad thing.

In terms of scope, we have tested only a limited range of models, and used only the audio-visual simultaneity-judgment task with austere stimuli. As noted above, a variety of blended or modified models could be entertained. Furthermore, there is at least one recently advocated class of model relevant to simultaneity judgments that we have ignored: Population-code (sometimes called labelled-line) models (Roach et al., 2011; Roseboom et al., 2015; Yarrow et al., 2015). However, there were reasons for leaving this class of model out. In the absence of some manipulation based on sensory adaptation, its basic simultaneity-judgment prediction is very similar to AT-A-GLANCE, but without the noisy criteria aspect. However, to deal with established differences in slope for the two sides of the simultaneity function, one would need to add something like noisy criteria. This remains entirely within the spirit of a population-code model, as the population of neurones simply supplies an estimate of the represented quantity, in this case subjective SOA, and is agnostic with regard to further steps to formulate a binary decision. Indeed, population-code accounts of the simultaneity judgment are perhaps best viewed as a more fleshed-out representational stage within an independent channels / general-threshold framework (Yarrow & Arnold, 2016). To this extent, the current result can be viewed as supportive of a population code (plus noisy criteria) as much as of AT-A-GLANCE.

We have also focussed here exclusively on modelling the simultaneity-judgment task. Of course, more general models are typically preferable to models which explain only one particular phenomenon. It is possible to extend models like those we test here to simultaneously account for data from multiple tasks. One example is Diederich and Colonius' (2015) simultaneous account of temporal order judgments and the redundant-signal effects data via an extension of the ELA model.

However, such efforts have thus far focussed on applying a single model to several tasks. Comparing such extended variants across several models, like those we describe here, via simultaneous fits, represents an interesting avenue for future research.

Regarding the degree to which results here can be generalised to all people – we are limited in what we can say about our sample, beyond stating that it was certainly not random, and likely primarily both young and WEIRD (Western, Educated, Industrialised, Rich and Democratic). We suspect that the way in which humans make decisions about the simultaneity of flashes and beeps is fairly universal (or at least universally idiosyncratic) but this is ultimately an empirical question for future research.

Conclusion

Here, we have demonstrated how to investigate experimental questions addressed using the simultaneity-judgment task by fitting Bayesian multilevel models, illustrating this approach with three recently advocated observer models. While the ELA and MCD models have some attractive features, for now we recommend researchers interested in this kind of approach consider using a model akin to AT-A-GLANCE, because the ultimate arbitrator between theories should probably be how well they predict out of sample data, and AT-A-GLANCE performed best in this regard. We have also shown that performance on the simultaneity-judgment task reflects an interpretation by the participant based on malleable decision criteria. It is these criteria that determine the width of the simultaneity function, and hence the window of subjective simultaneity. Thus, because of its strategic nature, this window casts only a thin light on multisensory temporal integration/binding processes, and should be interpreted with caution. Although no universal remedy, changes in measures that directly assess internal noise seem more pertinent when drawing conclusions about the causes underlying perceptual differences between clinical and other groups.

References

Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & Van Der Sluis, S. (2014). A solution to dependency: using multilevel analysis to accommodate nested data. *Nature Neuroscience, 17*(4), 491-496.

Alcalá-Quintana, R., & García-Pérez, M. A. (2013). Fitting model-based psychometric functions to simultaneity and temporal-order judgment data: MATLAB and R routines. *Behavior Research Methods, 45*(4), 972-998.

Allan, L. G. (1975a). The relationship between judgments of successiveness and judgments of order. *Perception and Psychophysics, 18*(1), 29-36.

Allan, L. G. (1975b). Second guesses and the attention-switching model for successiveness discrimination. *Perception & Psychophysics, 17*, 65–68.

Baron, J. (1971). The threshold for successiveness. *Perception & Psychophysics, 10*(4), 201-207.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bausenhart, K. M., Di Luca, M., & Ulrich, R. (2018). Assessing duration discrimination: Psychophysical methods and psychometric function analysis. *Timing and Time Perception: Procedures, Measures, & Applications* (pp. 52-78). Brill.

Bonnet, E., Masson, G. S., & Desantis, A. (2022). What over When in causal agency: Causal experience prioritizes outcome prediction over temporal priority. *Consciousness and Cognition, 104*, 103378. https://doi.org/10.1016/j.concog.2022.103378

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

Burr, D., Silva, O., Cicchini, G. M., Banks, M. S., & Morrone, M. C. (2009). Temporal mechanisms of multimodal binding. *Proceedings of the Royal Society B: Biologial Sciences, 276*, 1761-1769.

Cass, J., & Van der Burg, E. (2014). Remote temporal camouflage: Contextual flicker disrupts perceived visual temporal order. *Vision Research, 103*, 92-100.

Chen, Y., Lewis, T. L., Shore, D. I., & Maurer, D. (2017). Early Binocular Input Is Critical for Development of Audiovisual but Not Visuotactile Simultaneity Perception. *Current Biology, 27*(4), 583-589.

Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-window-of-integration model. *Journal of Cognitive Neuroscience, 16*(6), 1000-1009.

Diederich, A., & Colonius, H. (2015). The time window of multisensory integration: Relating reaction times and judgments of temporal order. *Psychological Review, 122*(2), 232.

Foucher, J. R., Lacambre, M., Pham, B., Giersch, A., & Elliott, M. A. (2007). Low time resolution in schizophrenia: lengthened windows of simultaneity for visual, auditory and bimodal stimuli. *Schizophrenia Research, 97*(1), 118-127.

Freeman, E. D., Ipser, A., Palmbaha, A., Paunoiu, D., Brown, P., Lambert, C., Leff, A., & Driver, J. (2013). Sight and sound out of synch: Fragmentation and renormalisation of audiovisual integration and subjective timing. *Cortex, 49*(10), 2875-2887.

Fujisaki, W., & Nishida, S. (2007). Feature-based processing of audio-visual synchrony perception revealed by random pulse trains. *Vision Research, 47*, 1075-1093.

García-Pérez, M. A., & Alcalá-Quintana, R. (2012a). On the discrepant results in synchrony judgment and temporal-order judgment tasks: a quantitative model. *Psychonomic Bulletin & Review, 19*(5), 820-846.

García-Pérez, M. A., & Alcalá-Quintana, R. (2012b). Response errors explain the failure of independent-channels models of perception of temporal order. *Frontiers in Psychology, 3*, 94. 10.3389/fpsyg.2012.00094

García-Pérez, M. A. (2014). Does time ever fly or slow down? The difficult interpretation of psychophysical data on time perception. *Frontiers in Human Neuroscience, 8*, 415.

García-Pérez, M. A., & Alcalá-Quintana, R. (2015). Converging evidence that common timing processes underlie temporal-order and simultaneity judgments: A model-based analysis. *Attention, Perception, & Psychophysics, 77*(5), 1750-1766.

García-Pérez, M. A., & Peli, E. (2014). The bisection point across variants of the task. *Attention, Perception, & Psychophysics, 76*(6), 1671-1697.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*(4), 455-467.

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

Habets, B., Bruns, P., & Röder, B. (2017). Experience with crossmodal statistics reduces the sensitivity for audio-visual temporal asynchrony. *Scientific Reports, 7*(1), 1-7.

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. *Journal of Experimental Psychology, 63*(3), 289.

Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: age-related changes in multisensory temporal processing. *Neuropsychologia, 49*(3), 461-467.

Holmes, N. P., & Spence, C. (2005). Multisensory integration: space, time and superadditivity. *Current Biology, 15*, R762-R764.

Horsfall, R., Wuerger, S., & Meyer, G. (2021). Visual intensity-dependent response latencies predict perceived audio–visual simultaneity. *Journal of Mathematical Psychology, 100*, 102471.

Johnston, P. R., Alain, C., & McIntosh, A. R. (2022). Individual Differences in Multisensory Processing Are Related to Broad Differences in the Balance of Local versus Distributed Information. *Journal of cognitive neuroscience*, 34(5), 846–863. https://doi.org/10.1162/jocn_a_01835

Lambert, B. (2018). *A student's guide to Bayesian statistics*. Sage.

Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds signals from multiple senses. *Proceedings of the National Academy of Sciences of the United States of America, 108*(51), 1441. 10.1073/pnas.1115267108 [doi]

Love, S. A., Petrini, K., Cheng, A., & Pollick, F. E. (2013). A psychophysical investigation of differences between synchrony and temporal order judgments. *PloS One, 8*(1), e54798.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide* (2nd ed.). Lawrence Erlbaum Associates.

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. *Frontiers in Psychology, 4:* 798*.*

Marsicano, G., Cerpelloni, F., Melcher, D. & Ronconi, L. (2022). Lower multisensory temporal acuity in individuals with high schizotypal traits: a web-based study. Scientific Reports, 12, 2782. https://doi.org/10.1038/s41598-022-06503-1

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *Journal of Neuroscience, 7*(10), 3215-3229.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology, 14*(2), 247-279.

Mioli, A., Diolaiuti, F., Zangrandi, A., Orsini, P., Sebastiani, L., & Santarcangelo, E. L. (2021). Multisensory Integration Is Modulated by Hypnotizability. *The International journal of clinical and experimental hypnosis, 69*(2), 215–224. https://doi.org/10.1080/00207144.2021.1877089

Mollon, J. D., & Perkins, A. J. (1996). Errors of judgement at Greenwich in 1796. *Nature, 380*, 101-102.

Morgan, M. J., Melmoth, D., & Solomon, J. A. (2013). Linking hypotheses underlying Class A and Class B methods. *Visual Neuroscience, 30*(5-6), 197-206.

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: the generalized linear mixed model. *Journal of Vision, 12*(11), 10.1167/12.11.26. 10.1167/12.11.26 [doi]

Navarra, J., & Fernández-Prieto, I. (2020). Perceptual association enhances intersensory temporal precision. *Cognition, 194*, 104089.

Noel, J., De Niear, M. A., Stevenson, R., Alais, D., & Wallace, M. T. (2017). Atypical rapid audio-visual temporal recalibration in autism spectrum disorders. *Autism Research, 10*(1), 121-129.

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. *Nature Communications, 7*, 11543.

Pesnot Lerousseau, J., Parise, C. V., Ernst, M. O., & van Wassenhove, V. (2022). Multisensory correlation computations in the human brain identified by a time-resolved encoding model. *Nature Communications, 13*, 2489.

Prins, N., & Kingdom, F. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes toolbox. *Frontiers in Psychology, 9*, 1250.

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences, 24*, 574-590.

Roach, N. W., Heron, J., Whitaker, D., & McGraw, P. V. (2011). Asynchrony adaptation reveals neural population code for audio-visual timing. *Proceedings of the Royal Society B: Biological Sciences, 278*(1710), 1314-1322. 10.1098/rspb.2010.1737

Roseboom W. (2019). Serial dependence in timing perception. Journal of experimental psychology. Human perception and performance, 45(1), 100–110. https://doi.org/10.1037/xhp0000591

Roseboom, W., & Arnold, D. H. (2011). Twice upon a time: multiple concurrent temporal recalibrations of audiovisual speech. *Psychological Science, 22*(7), 872-877. 10.1177/0956797611413293

Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision, 9*(12), 4.

Roseboom, W., Linares, D., & Nishida, S. (2015). Sensory adaptation for timing perception. *Proceedings of the Royal Society of London B: Biological Sciences, 282*(1805), 20142833.

Scarpina, F., Migliorati, D., Marzullo, P., Mauro, A., Scacchi, M., & Costantini, M. (2016). Altered multisensory temporal integration in obesity. *Scientific Reports, 6*, 28382. 10.1038/srep28382 [doi]

Schneider, K. A., & Bavelier, D. (2003). Components of visual prior entry. *Cognitive Psychology, 47*(4), 333-366.

Schwarz, W. (2006). On the relationship between the redundant signals effect and temporal order judgments: parametric data and a new model. *Journal of Experimental Psychology: Human Perception and Performance, 32*(3), 558.

Stan Development Team (2020). RStan: the R interface to Stan. http://mc-stan.org/

Stan Development Team (2022). Stan Modeling Language Users Guide and Reference Manual, version 2.30. https://mc-stan.org

Stelmach, L. B., & Herdman, C. M. (1991). Directed attention and perception of temporal order. *Journal of Experimental Psychology: Human Perception and Performance, 17*(2), 539-550.

Sternberg, S., & Knoll, R. L. (1973). The perception of temporal order: Fundamental issues and a general model. In S. Kornblum (Ed.), *Attention and Performance IV* (pp. 629-686). Academic Press.

Stevenson, R. A., Baum, S. H., Krueger, J., Newhouse, P. A., & Wallace, M. T. (2018). Links between temporal acuity and multisensory integration across life span. *Journal of Experimental Psychology: Human Perception and Performance, 44*(1), 106.

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., & Wallace, M. T. (2014). The impact of multisensory integration deficits on speech perception in children with autism spectrum disorders. *Frontiers in Psychology, 5*, 379.

Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., & Porter, N. R. (2002). When is Now? Perception of simultaneity. *Proceedings of the Royal Society of London Series B: Biological Sciences, 268*, 31-38.

Stroud, J. M. (1956). The fine structure of psychological time. In H. Quastler (Ed.), *Information theory in Psychology* (pp. 174-205). Free Press.

Swets, J., Tanner, W. P., Jr. & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review, 68,* 301-40.

Ulrich, R. (1987). Threshold models of temporal-order judgments evaluated by a ternary response task. *Perception and Psychophysics., 42*(3), 224-239.

van Eijk, R. L., Kohlrausch, A., Juola, J. F., & van de Par, S. (2008). Audiovisual synchrony and temporal order judgments: effects of experimental method and stimulus type. *Perception and Psychophysics., 70*(6), 955-968.

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413-1432.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., & Bürkner, P. (2021). Rank-normalization, folding, and localization: an improved R for assessing convergence of MCMC (with discussion). *Bayesian Analysis, 16*(2), 667-718.

Venables, P. H. (1960). Periodicity in reaction time. *British Journal of Psychology, 51*, 37-43.

Vroomen, J., & Keetels, M. (2020). Perception of causality and synchrony dissociate in the audiovisual bounce-inducing effect (ABE). *Cognition, 204*, 104340.

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

Yarrow, K., Jahn, N., Durant, S., & Arnold, D. H. (2011). Shifts of criteria or neural timing? The assumptions underlying timing perception studies. *Consciousness and Cognition, 20*, 1518-1531. 10.1016/j.concog.2011.07.003

Yarrow, K. (2018). Collecting and interpreting judgments about perceived simultaneity: A model-fitting tutorial. In A. Vatakis, F. Balci, M. Di Luca, Á. Correa (Eds.), *Timing and Time Perception: Procedures, Measures, & Applications* (pp. 295-325). Brill.

Yarrow, K., & Arnold, D. H. (2016). The Timing of Experiences: How Far Can We Get with Simple Brain Time Models? In B. Mölder, V. Arstila & P. Øhrstrøm (Eds.), *Philosophy and Psychology of Time* (pp. 187-201). Springer.

Yarrow, K., Kohl, C., Segasby, T., Bansal, R. K., Rowe, P., & Arnold, D. H. (2022). Neural-latency noise places limits on human sensitivity to the timing of events. *Cognition, 222*, 105012.

Yarrow, K., Martin, S. E., Di Costa, S., Solomon, J. A., & Arnold, D. H. (2016). A roving dual-presentation simultaneity-judgment task to estimate the point of subjective simultaneity. *Frontiers in Psychology, 7,* 416.

Yarrow, K., Minaei, S., & Arnold, D. H. (2015). A model-based comparison of three theories of audiovisual temporal recalibration. *Cognitive Psychology, 83*, 54-76.

Yarrow, K., & Roseboom, W. (2017, October 31). Should multisensory temporal acuity be viewed through the window of perceived simultaneity? https://doi.org/10.31234/osf.io/zpkq7

Yarrow, K., Sverdrup-Stueland, I., Roseboom, W., & Arnold, D. H. (2013). Sensorimotor Temporal Recalibration Within and Across Limbs. *Journal of Experimental Psychology: Human Perception and Performance, 39*(6), 1678-1689.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. *Perception & Psychophysics, 67*(3), 531-544.

Appendix A: Multilevel model specifications

AT-A-GLANCE model implementation

Single-level AT-A-GLANCE

Our first multilevel model built upon the AT-A-GLANCE four-parameter single-level observer model (Yarrow et al., 2011). Our description of that model here is more complete than in any of our previous papers and thus supersedes them. Under this account, the observer judges two stimuli simultaneous when the internal signals they generate arrive at a decision centre with a subjective SOA that is both above a (noisy) low criterion and below a (noisy) high criterion. Hence AT-A-GLANCE implies three normally distributed random variables: Two decision criteria (c_{L} and c_{H}) used to demarcate successive judgments from simultaneous judgments, and the subjective SOA, s. These three random variables can be expressed as a single, trivariate normal random variable, with mean and variance:

(A1) $$\boldsymbol{\mu} = \begin{pmatrix} \mu_{\mathrm{L}} \\ \mu_{\mathrm{S}} \\ \mu_{\mathrm{H}} \end{pmatrix}, \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{\mathrm{L}}^2 & \rho_{\mathrm{LS}}\sigma_{\mathrm{L}}\sigma_{\mathrm{S}} & \rho_{\mathrm{LH}}\sigma_{\mathrm{L}}\sigma_{\mathrm{H}} \\ \rho_{\mathrm{LS}}\sigma_{\mathrm{L}}\sigma_{\mathrm{S}} & \sigma_{\mathrm{S}}^2 & \rho_{\mathrm{SH}}\sigma_{\mathrm{S}}\sigma_{\mathrm{H}} \\ \rho_{\mathrm{LH}}\sigma_{\mathrm{L}}\sigma_{\mathrm{H}} & \rho_{\mathrm{SH}}\sigma_{\mathrm{S}}\sigma_{\mathrm{H}} & \sigma_{\mathrm{H}}^2 \end{pmatrix}$$

Let $f(c_{\mathrm{L}}, s, c_{\mathrm{H}})$ denote its density. Then:

(A2) $$P(C_{\mathrm{L}} < S < C_{\mathrm{H}}) = \int_{-\infty}^{\infty} dc_{\mathrm{H}} \int_{-\infty}^{c_{\mathrm{H}}} ds \int_{-\infty}^{s} dc_{\mathrm{L}}\, f(c_{\mathrm{L}}, s, c_{\mathrm{H}})$$

Unfortunately, expressed in this way the (single-level) model has a hefty eight parameters (excluding μ_{S}, which equals the experimental SOA). We can easily take a view regarding the ρ parameters, for example fix them to 0 for fully uncorrelated sources of noise, but Equation A1 is still slow to evaluate (we are not aware of a closed-form solution) and likely degenerate with regard to the three σ parameters (meaning they can trade off against each other to give near identical predictions). However, if we assume ρ_{LH} = 1 and ρ_{LS} = ρ_{SH} = 0, implying correlated noise in the two criteria, a closed-form approximation is available:

(A3) $P(S|\Delta t) \approx \Phi\left(\frac{\Delta t - c_{\mathrm{L}}}{\sigma_{\mathrm{L}}}\right) - \Phi\left(\frac{\Delta t - c_{\mathrm{H}}}{\sigma_{\mathrm{H}}}\right)$

where S denotes the event that the observer responds “simultaneous”, Δt is the SOA, and Φ is the standard normal cumulative distribution function. The σ values quantify (inversely) the slope on each side of the psychometric function. These are composite noise variables, used because they are formally identifiable in a model fit (meaning that they do not trade off perfectly) whereas the various psychological constructs that feed into them are not. Each σ, when squared, represents the sum of two sources of variance. The first, the variance of subjective SOAs (σ_{S}^2 from Equation A1) is itself derived from the (Gaussian) latency variance associated with each stimulus (if we assume uncorrelated sensory channels, it is their sum). This source contributes to the slope on both sides of the psychometric function (low and high). The second, the trial-by-trial (Gaussian) variance in a decision criterion (σ_{L}^2 or σ_{H}^2 from Equation A1) has a unique magnitude on each side of the function, thus allowing the slopes to vary. Because Equation A3 is an approximation, we used a slower-to-evaluate method when the approximation breaks down – this can be seen in our code as an “override” function.[10]

To this model, we first added a lapse parameter, l, such that participants are assumed to make an (unbiased) guess on a proportion of trials equalling $2l$. This effectively forms upper and lower limits on the psychometric function at l and 1- l. We also opted to reparametrize c_{L} and c_{H} in terms of their midpoint and the distance between them.[11] We call these τ and $\Delta\delta$ respectively, to make the terminology comparable with near-equivalent parameters from a second observer model, ELA,

---

[10] Equation A3 breaks down when the difference between σ_{L} and σ_{H} is large relative to the distance between c_{H} and c_{L}. Our override function is essentially a simulation, rather than a numerical implementation of Equation A2. However, because Stan requires deterministic predictions, and to reduce computation time, in place of random sampling for each source of noise, we divided the probability space from .01 to .99 in 50 steps of .02, and applied an inverse Gaussian function to these values to recover pseudo-simulated noise scores. This process can distort model predictions slightly relative to a true Monte Carlo simulation, but informal explorations suggested this distortion was negligible.

[11] Reparameterization is often helpful in Stan programming to make the posterior likelihood surface more amenable to sampling, in part by allowing us to make better use of sensible hard constraints (and soft priors) on the values that (sampled) parameters can take. However, if needed or desired, the original (often more intuitive) parameters can easily be calculated using a ”transformed parameters” code block, something we included in our code and analysis here.

described later. We also reparametrized σ_{H} by instead sampling the posterior based on m, the natural log of a quantity applied as a multiplier to σ_{L} in order to determine σ_{H}:

(A4) $$\sigma_{\mathrm{H}} = \exp(m)\sigma_{\mathrm{L}}$$

Hence the reparametrized single-level model incorporating lapsing becomes:

(A5) $$P(S|\Delta t) \approx l + (1-l)\left[\Phi\left(\frac{\Delta t - \tau + \Delta\delta/2}{\sigma_{\mathrm{L}}}\right) - \Phi\left(\frac{\Delta t - \tau - \Delta\delta/2}{\exp[m]\sigma_{\mathrm{L}}}\right)\right]$$

For completeness, we next implemented priors (and provide accompanying code) for a single-level Bayesian implementation of this model, but here move straight to describing the multilevel case, which estimates the abovementioned five parameters for each of our 19 participants at once.

Multilevel AT-A-GLANCE, one condition.

Moving to a multilevel model requires moving from a scalar to a vector of parameters for each of the participant-level parameters already described. However, on its own this only gives us a "heterogeneous" model. A full "hierarchical" or multilevel model also requires the addition of group-level parameters (to capture random variation in participant-level parameters across the group) and, in the case of Bayesian models, "hyperpriors" (meaning expectations regarding sensible values for the group-level parameters based on what is known before the current data are collected).

To the 95 participant-level parameters (coded as five vectors/arrays) we therefore added a set of group-level parameters. Multilevel models require that we specify a distribution (for example normal) which describes the way each participant-level parameter varies across the group. The parameters of these distributions are then estimated alongside the individual-level parameters: In effect, when determining the likelihood of a set of parameters for a particular participant, we consider both the likelihood of their data given their participant-level parameters, and the likelihood of those participant-level parameters given the group-level distribution from which they are presumed to be being drawn.

For the AT-A-GLANCE model, we specified a normal group-level distribution for parameter τ, the midpoint of the two criteria for judging a stimulus simultaneous. The normal distribution is a good default choice for unbounded continuous parameters, and conforms to what would be assumed by a second-stage procedure such as applying a t-test to individual parameter estimates (a choice that is generally well justified by the central limit theorem). We included both the mean (μ_τ) and the standard deviation (σ_τ) of this distribution as parameters for estimation within the model.[12]

For each group-level distribution parameter, Bayesian modelling encourages us to also specify a (hyper)prior distribution, based on our subject-specific knowledge. This is a somewhat uncomfortable step for those with a frequentist background, but hyperpriors can be made as uninformative/diffuse as the modeller desires (at least when considering just the untransformed parameter). Furthermore, the alternative perspective is quite hard to defend. It implies that any and all values for a group-level summary statistic such as the mean midpoint of perceptual synchrony are equally likely before we see our particular set of data. However, using μ_τ as an example, even in a case study of a patient with a specific relative-timing related pathological deficit, the reported point of subjective simultaneity was only +210 ms (Freeman et al., 2013). Hundreds of group averages of similar measures have been reported in the literature, and although we have not reviewed them all, we are confident that *all* are much closer to zero than to, say, ±1000 ms.

Here, we utilised an (extremely diffuse) Cauchy hyperprior on μ_τ, with location of 0 and scale of 800 ms. Our code defaults to setting the former to the (unweighted) mean SOA in the data set and the latter to the range of asynchronies used, but the user can override these and several other hyperprior choices via parameters passed to Stan from R as part of the data set. For the hyperprior on σ_τ, which should be zero-bounded, we used a lognormal distribution with μ of 5.59

[12] In moving to a multilevel model, we applied what is known as a "non-centred" parameterisation to some group-level parameters in order to try and reduce correlations between group-level and participant-level parameters (see the Stan manuals for further details). This approach was applied for τ and also the β coefficients representing changes across conditions (described later). Essentially, we modelled variation across the group using a standard normal distribution, then derived scaled values of τ for each participant by multiplying this standardised variation by the group σ before adding the group μ.

and σ of 1. The code defaults μ to the natural log of one-third the range of asynchronies in the data, which, along with an σ of 1, for our data gives a right-skewed distribution with a mode of ≈ 100 ms. Note that the μ parameter of a lognormal distribution is not in fact its mean, which is instead obtained as $\exp\left(\mu + \frac{\sigma^2}{2}\right)$. Hence applying this transformation is sensible when subsequently interpreting parameters of this kind. In sum – we expected τ to be normally distributed across the group, with a group mean vaguely near zero ms and a group SD vaguely near 100 ms.

For $\Delta\delta$, the distance between the two judgment criteria, which is zero-bounded, we specified a lognormal group-level distribution and had the model estimate both parameters (μ_{δ} and σ_{δ}). For hyperpriors on μ_{δ} and σ_{δ}, we used normal and lognormal distributions respectively, the former with a μ of 5.59 and σ of 1.4 and the latter with μ of 1.1 and σ of 1 (the code again defaults to basing some of these on the range of asynchronies found in the data). This translates to expecting $\Delta\delta$ to vary across the group according to heavily right-skewed distribution with a mode vaguely near 90 ms, but with hyperpriors giving plenty of scope for very different central tendencies and shapes to emerge.

For σ_{L}, the inverse slope of the left side of the psychometric function, we specified a lognormal group-level distribution (and hyperpriors on its two parameters, μ_{σ} and σ_{σ}) in exactly the same way as outlined above for $\Delta\delta$.

For m, a parameter which is used to create σ_{H} by multiplicatively modifying σ_{L} (see Equation A4), we specified a normal group-level distribution and estimated both the mean (μ_{m}) and the standard deviation (σ_{m}). Because of the exponentiation in Equation A4, values of m below zero lead to $\sigma_{\mathrm{H}} < \sigma_{\mathrm{L}}$, and vice versa for values above 0. Hence, we placed a normal hyperprior on μ_{m} with a mean of zero. We sought to prevent the fit from favouring extreme differences in slope on the two sides of the function, as this is against the spirit of the model, which posits a substantial source of shared noise affecting both sides. Any difference comes from criterial noise that, if too extreme, would imply regular illogical positioning ($\mathrm{C_L} > \mathrm{C_H}$) on individual trials. Hence, we gave this hyperprior an SD of 0.5 (which has the effect of making identical slopes around 11 times as likely, a priori, as slopes that differ by a factor of 3). For σ_{m} we used a lognormal hyperprior with μ of -0.69 and σ of 1 (equating to an expectation of group SD vaguely near 0.2).

Finally, for l, the parameter capturing lapses of attention, we specified a beta group-level distribution, as these deal well with parameters that are 0-1 bounded such as proportions. Beta distributions are defined by two parameters, but we wanted to keep our model simple and also place strong expectations for a lapse rate near zero. We therefore fixed the second parameter, β_{l}, to 50, and estimated only the group's modal guess rate (θ_{l}) which determined the first beta-distribution parameter, α_{l}, according to:

(A6) $$\alpha_{\mathrm{l}} = \frac{2\theta_{\mathrm{l}} - \theta_{\mathrm{l}}\beta_{\mathrm{l}} - 1}{\theta_{\mathrm{l}} - 1}$$

We used a beta hyperprior on θ_{l} with α of 1.49 and β of 50. This equates to strongly expecting a group modal lapse rate around 1%.

Multilevel AT-A-GLANCE, differences across conditions.

Up to this point we have described a multilevel AT-A-GLANCE model with 104 parameters, capable of describing simultaneity-judgment data from 19 participants in a single experimental condition. We include accompanying code for this model so readers can see the additions required to go from 1) single-level, to 2) single-condition multilevel, to 3) multi-condition multilevel model, which is our final destination. To get to this final model, we still need to specify additional parameters describing how one or more of our participant-level parameters can vary across conditions of the experiment. We also need to update our model predictions to incorporate the effects of these parameters. As noted in the main text methods, this last set of parameters are conceptually akin to regression coefficients, affecting the model prediction contingent on the value of the conservative and rebound dummy codes. Dummy codes are 0 or 1 values denoting membership of a particular condition, included as columns within long-form data, where the dependent variable appears in a single column and other columns carry information about participant, condition and so forth.

The AT-A-GLANCE model envisages participants utilising two criteria to interpret a subjective difference in arrival times as simultaneous or not. Hence, instructions to be more conservative can be dealt with by allowing these two criteria to move. However, as previously described, we reparametrized the criteria as τ, their midpoint, and $\Delta\delta$, their difference, so it is these parameters that should be allowed to vary. Each participant therefore required a set of coefficients, $\beta_{\mathrm{\tau C}}$, $\beta_{\mathrm{\tau R}}$, $\beta_{\mathrm{\delta C}}$, and $\beta_{\mathrm{\delta R}}$, to represent change (compared to baseline). The first subscript represents the parameter being adjusted and the second represents the Conservative and Rebound conditions. However, we were mindful that while τ is unbounded, $\Delta\delta$ has a zero lower bound. Hence we allowed straightforward additive changes to τ, but only positive multiplicative ones to $\Delta\delta$, with the latter implemented by exponentiating the relevant coefficient such that positive/negative values translate to multiplication by greater than or less than 1 respectively. This yields the heterogenous model of Eqns. 1 and 2 (see main text).

All that now remains to be done for this model is to describe the estimation of group-level distributions for the experimental effects (the four β coefficients), along with the associated hyperpriors. For each of these coefficients we specified a normal group-level distribution and estimated both mean ($\mu_{\ldots}$) and standard deviation ($\sigma_{\ldots}$) parameters (implying eight further group-level parameters). The parameters of these group-level distributions essentially mirror the terms commonly described as “fixed” and “random” effects (respectively) within a frequentist general(ised) linear multilevel model framework: The former describe the group-mean effects, the latter the variation in these effects across participants. We constrain their values with normal hyperpriors (which due to zero bounding are effectively half-normal for $\sigma_{\ldots}$ parameters) with μs of 0 and σs of either 80 (for $\mu_{\mathrm{\tau}\ldots}$ and $\sigma_{\mathrm{\tau}\ldots}$ hyperpriors) or 1 (for $\mu_{\mathrm{\delta}\ldots}$ and $\sigma_{\mathrm{\delta}\ldots}$ hyperpriors). To summarise – we expected zero-size mean effects with zero SD across the group, but modest and even fairly large effects (and associated variation in effects) would not be unexpected. The final model has 188 parameters (five core plus four β parameters for each of 19 participants, plus nine parameters describing group variation in core parameters and eight parameters describing group variation in β parameters). These were estimated based on 513 data points (19 participants x 3 conditions x 9 SOAs).

Multilevel AT-A-GLANCE's alternative account for conservative behaviour

The model described so far can fit simultaneity-judgment data from multiple participants at once and capture changes across conditions in terms of an adjustment of parameters quantifying decision criteria. This model essentially represents the hypothesis that simultaneity judgments are subject to strategic alteration based on these decision criteria. We also created an alternative *hard-threshold* model variant, in which participants are assumed to maintain their threshold from the pre-test but, in the "be conservative" condition, respond "synchronous" on a random subset of trials in which they actually perceive synchrony. This model essentially represents the hypothesis that what participants initially report in a simultaneity-judgment task is determined by a structural thresholding mechanism that does not yield to their subsequent strategic imperatives. This might be the same gating mechanism underlying multisensory binding/integration if that type of process is also viewed as all-or-none from a temporal perspective.

The hard-threshold multilevel AT-A-GLANCE model we applied is identical to the multilevel AT-A-GLANCE model described so far, except in relation to the set of β coefficients used to permit changes across conditions. Instead of allowing changes to two criteria (in each of two conditions, relative to the baseline), we now utilise just one change per condition – a proportional reduction in the number of trials judged synchronous. This can be represented by a pair of coefficients, β_{C} and β_{R}, and yields a heterogenous foundation with a binomially distributed number of "simultaneous" responses:

(A7) $S_{X\Delta t} \sim B(30, \beta_X[l + p_{B\Delta t} - lp_{B\Delta t}])$, where $X \in \{B, C, R\}$ and $p_{B\Delta t}$ is defined in Equation 2 (main text, Results).

For the β_{C} and β_{R} parameters, we modelled variation at the group level as a beta distribution, but parameterised in terms of a mean parameter:

(A8) $$\varphi_{\dots} = \frac{\alpha}{\alpha+\beta}$$

And a total count parameter:

(A9) $$\lambda_{\dots} = \alpha + \beta$$

We followed recommendations in the Stan documentation by specifying hyperpriors that were beta (α = 1, β = 1, implying uniform) and pareto ($y_{\min}$ = 0.1, α = 0.5) for $\varphi_{\dots}$ and $\lambda_{\dots}$ respectively.

ELA model implementation

Single-level ELA

Our second class of multilevel model built on the four-parameter ELA model (García-Pérez & Alcalá-Quintana, 2012) which predicts reports of simultaneity as the integral (between two decision boundaries) of a difference of exponential distributions. This prediction is described by:

(A10) $$P(S|\Delta t) = F(\Delta\delta/2\,;\Delta t) - F(-\,\Delta\delta/2\,;\Delta t),$$

where function F is given by:

(A11) $$F(d;\Delta t) = \begin{cases} \frac{\lambda_{\mathrm{a}}}{\lambda_{\mathrm{a}}+\lambda_{\mathrm{v}}} \exp[-\lambda_v(\Delta t - \tau + d)] & \text{if } d \le \Delta t - \tau \\ 1 - \frac{\lambda_{\mathrm{v}}}{\lambda_{\mathrm{a}}+\lambda_{\mathrm{v}}} \exp[\lambda_a(\Delta t - \tau + d)] & \text{if } d > \Delta t - \tau \end{cases}.$$

Under this model, λ_{a} and λ_{v} are the rate parameters of (shifted) exponential distributions of arrival times (at the decision centre) for the auditory and visual signals respectively. We have reversed the sign of García-Pérez and Alcalá-Quintana's τ parameter, making it directly comparable to AT-A-

GLANCE's midpoint between two judgment criteria used to categorise subjective asynchronies as simultaneous. Otherwise, our Equation A11 is identical to their Eqn. 3.

For even further ease of comparison with AT-A-GLANCE, we consider the inverse of the λ_{a} parameter ($\lambda_{\mathrm{a}}^{-1}$), whose values have a scale and meaning similar to those of AT-A-GLANCE's two noise parameters. Hence higher values equate to a higher level of internal noise and reduced precision. Furthermore, in place of λ_{v} we sampled for m, the natural log of a quantity applied as a multiplier to $\lambda_{\mathrm{a}}^{-1}$ in order to determine the inverse of λ_{v}, in a manner analogous to that described in Equation A4 above for AT-A-GLANCE's second noise parameter. Finally, we also included the same lapse parameter used in our implementation of AT-A-GLANCE, l, such that participants were assumed to make an (unbiased) guess on a proportion of trials equalling $2l$. This leads to the following prediction:

(A12) $$P(S|\Delta t) = l + (1-l)[F(\Delta\delta/2\,;\Delta t) - F(-\Delta\delta/2\,;\Delta t)],$$

where function F is given by:

(A13) $$F(d;\Delta t) = \begin{cases} \frac{\exp\left[m-(\Delta t-\tau-d)/\left(e^{m}\lambda_a^{-1}\right)\right]}{\exp(m)+1} & \text{if } d \le \Delta t-\tau \\ 1-\frac{\exp\left[m+(\Delta t-\tau-d)/\lambda_a^{-1}\right]}{\exp(2m)+\exp(m)} & \text{if } d > \Delta t-\tau \end{cases}.$$

Multilevel ELA

With both AT-A-GLANCE and ELA utilising a set of largely analogous single-level parameters, we were able to develop multilevel models of ELA in a very similar way to that outlined above for AT-A-GLANCE. Hence, we added μ_{τ} and σ_{τ} parameters to describe the normal group-level distribution of τ, with Cauchy and lognormal hyperpriors respectively. Similarly, for the lognormal group-level distribution of $\Delta\delta$, we introduced μ_{δ} and σ_{δ}, with normal and lognormal hyperpriors respectively, as per the same parameter's treatment in AT-A-GLANCE. The (lognormal) group-level $\lambda_{\mathrm{a}}^{-1}$ in ELA was dealt with just like the group-level σ_{L} from AT-A-GLANCE, by including $\mu_{\lambda\mathrm{a}}$ and $\sigma_{\lambda\mathrm{a}}$ parameters with normal and lognormal hyperpriors respectively. Similarly, we included μ_{m} and σ_{m} to describe the normal group-level distribution of m, with normal and lognormal hyperpriors respectively, while for l, we added θ_{l} to define the mode of a beta group-level distribution (with a beta hyperprior). Finally, we added eight parameters to model the means and SDs of the normal group-level distributions for the four β coefficients which describe changes to τ and δ across experimental conditions (for example $\mu_{\tau\mathrm{C}}$ and $\sigma_{\tau\mathrm{C}}$ for the participant-level parameter $\beta_{\tau\mathrm{C}}$ adjusting τ in the conservative condition). These were specified with normal hyperpriors. We also constructed an alternative model describing the hard-threshold account, with group-level beta distributions of the function multiplier coefficients β_{C} and β_{R}, each described by mean and total count parameters with beta and pareto hyperpriors respectively, in place of changes to τ and $\Delta\delta$. In all but a handful of cases, hyperpriors had parameters exactly as specified for the analogous case in AT-A-GLANCE. The key exceptions were μ_{δ} and $\mu_{\lambda\mathrm{a}}$, relating to the distance between criteria and noise for the auditory stimulus respectively, for which we specified a slightly lower expectation (via setting μ = 5.08, with this default based on 1/5th of the range of SOAs). In the case of μ_{δ}, this followed from a programming choice – we sampled for values of $\Delta\delta/2$ rather than $\Delta\delta$, and hence μ_{δ} should be around half as large of the equivalent parameter from AT-A-GLANCE. In the case of $\mu_{\lambda\mathrm{a}}$, estimates for this parameter from past research tend to be lower than those obtained for μ_{σ}.

MCD model implementation

Single-level MCD

Our final class of model was built upon a three-parameter SJ-only implementation of Parise and Ernst's (2016) MCD model. This model describes how time-varying visual and auditory signals ($S_{\mathrm{v}}(t)$, $S_{\mathrm{a}}(t)$) are transformed into a time-varying synchrony signal which can then be averaged over the period following stimulus presentation to yield perceived synchrony for that trial (MCD_{Corr}). This process requires three kinds of filter, two applied in an early stage and one at a later stage, but all of the following form:

(A14) $f_{\mathrm{mod}} = t\exp(-t/\tau_{\mathrm{mod}})$

Where f_{mod} is an early modality-dependent filter (f_{a} and f_{v}) or a late multisensory filter (f_{av}), and τ_{mod} is the corresponding filter time constant. The final synchrony estimate is essentially the time-averaged output formed by multiplying together signals from two units. Each unit multiplies a single (early) filtered version of one modality with a double (early+late) filtered version of the other. The final synchrony estimate is then:

(A15)

$$MCD_{\mathrm{Corr}} = \frac{1}{t_2 - t_1}\int_{t_1}^{t_2} \left([S_{\mathrm{v}}(t) * f_{\mathrm{v}}(t)] \cdot \{[S_{\mathrm{a}}(t) * f_{\mathrm{a}}(t)] * f_{\mathrm{av}}(t)\}\right) \cdot \left([S_{\mathrm{a}}(t) * f_{\mathrm{a}}(t)] \cdot \{[S_{\mathrm{v}}(t) * f_{\mathrm{v}}(t)] * f_{\mathrm{av}}(t)\}\right) dt,$$

where * denotes convolution and · denotes pointwise multiplication. Finally, MCD_{Corr} is used to form binary judgments about synchrony by setting a criterion, with either MCD_{Corr} itself or the position of the criterion assumed to be affected by Gaussian trial-by-trial noise, yielding the prediction:

(A16) $P(S|\Delta t) = \Phi\left(\frac{MCD_{\mathrm{Corr}} - C}{\sigma}\right)$

For our fits, we fixed τ_{av} to 786 ms and τ_{v} to 87 ms based on fits to other data sets (Parise & Ernst, 2016) and allowed τ_{a}, σ and C to vary for each observer. We calculated MCD_{Corr} across a 14 second window centred on the arrival time of the first stimulus (and set to zero except for 10 ms on-off pulses for each signal). We also normalised it by dividing it by the unnormalised MCD_{Corr} for a synchronous input (MCD_{CorrS}). This normalisation meant that C could be expected to lie in a range bounded by 0 and just over 1, and σ should be interpretable on a similar scale. Because Stan does not currently offer built-in functions for convolution or fast Fourier transformation, we first determined MCD_{Corr} for values of τ_a from 1 to 200 ms in R, then passed them to Stan as a lookup table. Within the Stan code, MCD_{Corr} values from this table were made continuous (and hence differentiable) via quadratic interpolation. We also added a lapse parameter, l, consistent with that applied in our other models:

(A17) $$P(S|\Delta t) = l + (1-l)\Phi\left[\frac{\left({MCD_{\mathrm{Corr}}}/{MCD_{\mathrm{CorrS}}}\right)-C}{\sigma}\right]$$

Multi-level MCD

To upgrade to a multilevel MCD model we dealt with parameter l as per our previous models, by adding θ_{l} to define the mode of a group-level beta distribution, and specifying a beta hyperprior on it. For filter time constant τ_{a}, we specified a lognormal group-level distribution and estimated its two parameters, $\mu_{\tau\mathrm{a}}$ and $\sigma_{\tau\mathrm{a}}$. For $\mu_{\tau a}$ we specified a normal hyperprior, with μ = 4.34 and σ = 1.09, while the hyperprior for $\sigma_{\tau\mathrm{a}}$ was lognormal with μ = -1.39 and σ = 0.25. Together these correspond to a modal expectation for τ_{a} of around 73 ms. This is comparable to the value of 68 ms obtained by Parise and Ernst (2016). For criterion C, we specified a normal group-level distribution and estimated its two parameters, μ_{C} and σ_{C}. We gave μ_{C} a normal hyperprior, with μ and σ both set at 0.5, and σ_{C} a lognormal hyperprior with μ = 0.41 and σ = 1, together implying that C should have a group mean around 0.5 and SD around 0.55. We then specified a lognormal group-level distribution for internal-noise parameter σ and estimated both of this distribution's parameters, μ_{σ} and σ_{σ}. We gave μ_{σ} a normal hyperprior, with μ and σ set at -0.69 and 1 respectively, and σ_{σ} a lognormal hyperprior with μ = 3 and σ = 1, together implying that σ was a priori expected to have a group mode around 0.18, but with larger values remaining plausible. Finally, to allow behaviour to change across conditions, we allowed criterion C to vary via the introduction of two participant-level β coefficients, β_{CC} and β_{CR}, with multiplicative adjustments to C determined using their exponents. Each had an associated normal group-level distribution for which we estimated both mean ($\mu_{\mathrm{C}\ldots}$) and standard deviation ($\sigma_{\mathrm{C}\ldots}$) parameters (implying four further group-level parameters). Hyperpriors on these parameters were normal (effectively half-normal for the zero-bounded $\sigma_{\mathrm{C}\ldots}$ parameters) with μ = 0 and σ = 1. As for our other models, an alternative hard-threshold account was also tested, in which β_{CC} and β_{CR} were replaced with the function multiplier coefficients β_{C} and β_{R}, each described by mean and total count parameters with beta and pareto hyperpriors respectively.

1498

Appendix B: Adequacy of likelihood surface recovery

Before we can consider whether a model is a good description of reality, we need to determine whether we have successfully explored/characterised the posterior likelihood of the model and its parameters given the data. A model may in principle be perfectly correct, but in practice be impossible to evaluate because of issues such as degeneracy, where parameters trade off so that several different combinations can provide a similarly good fit. Table B1 summarises, for the two variants of each of our three models, a set of posterior exploration diagnostics showing how successfully the HMC NUTS algorithm was able to characterise the posterior in each case.

Table B1. Posterior exploration diagnostics.

Model	**< Posterior exploration diagnostics >**			
	% Divergent iterations	**Max $\hat{R}$**	**Minimum bESS**	**Minimum tESS**
AT-A-GLANCE criterial	**0.017**	**1.045**	**154**	**425**
AT-A-GLANCE hard threshold	**0.013**	**1.002**	**2324**	**4322**
ELA criterial	**0.013**	**1.002**	**5161**	**6740**
ELA hard threshold	**0.010**	**1.002**	**3883**	**5874**
MCD criterial	**0**	**1.001**	**6794**	**11927**
MCD hard threshold	**0**	**1.002**	**5177**	**2098**

For the AT-A-GLANCE model variant that allowed changes in criteria across conditions, diagnostics did not completely meet recommendations (Vehtari et al., 2021) despite a relatively long fit time (around 24 hours per run). In particular, alongside a very small percentage of divergent iterations, not all parameters reached the ideal level of mixing between chains ($\hat{R}$ < 1.01) or the suggested bulk and tail effective sample sizes (bESS and tESS >400). However, the vast majority of parameters did meet these recommendations. Furthermore, for the worst offending parameter (μ_τ), despite a bESS of only 154 the resulting Monte-Carlo standard error (a measure of the precision of parameter estimation) was just 0.95 (in the context of a mean value of 32.4 ms). The model predictions also mapped well onto the data (see main text Results). We therefore believe that from a practical point of view, this model was characterised adequately to allow us to make sensible comments regarding how well it described the data compared to other models explored here.

For the second, hard-threshold, variant of the AT-A-GLANCE model, exploration diagnostics met all recommendations with the exception of a very small percentage of divergent iterations. The posterior exploration diagnostics from Table 1 also indicate that both variants of both ELA and MCD models met recommendations in terms of chain mixing and bulk and tail effective sample sizes, with only a very small percentage of divergent iterations (<=0.013%). This suggests that the HMC NUTS sampling algorithm was able to properly characterise the posterior in each case. The ELA model's posterior proved particularly easy to characterise, with fits returning in under 30 minutes for these data. Parameter recovery was assessed separately (via simulation) – see Appendix D.

Appendix C: Assessment of Bayesian design choices

In implementing Bayesian multilevel models, we had to make various decisions, including specifying the distributions with which we expected participant-level parameters to vary across the group. We also had to set prior expectations for the parameters of these group-level distributions, known as hyperpriors. In Figure C1, we consider these choices for the AT-A-GLANCE model variant permitting changes in decision criteria across conditions. The smaller graphs within Figure C1 illustrate the posterior distributions obtained. They focus on the subset of group-level parameters relating to behaviour in the baseline condition (parts a-e). We also illustrate two of the remaining eight group-level parameters that relate to changes in behaviour in the conservative condition (specifically changes in the width of the simultaneity function; part f). Hyperpriors (plotted only across the limited range required to characterise the posteriors) are shown for comparison (dashed lines). Posteriors are markedly less diffuse than hyperpriors, rarely coincide with their modes, and don't appear to have been constrained by their edges. It is thus clear that posteriors were not unduly influenced by our choices regarding hyperpriors, and must have been very largely determined by the data. The figure also illustrates how these group-level estimates in turn parameterise group-distributions (shown as black against red lines in larger graphs). These describe how participant-level parameters vary across the group. They can be compared with the model's participant-level estimates for each individual (shown as circles) and a kernel-density plot derived from them (shown opposite parametric predictions). In general, the choices of distributions seem reasonable, although the participant-level estimates will have been constrained by these choices such that we are in part assessing a self-fulfilling prediction.

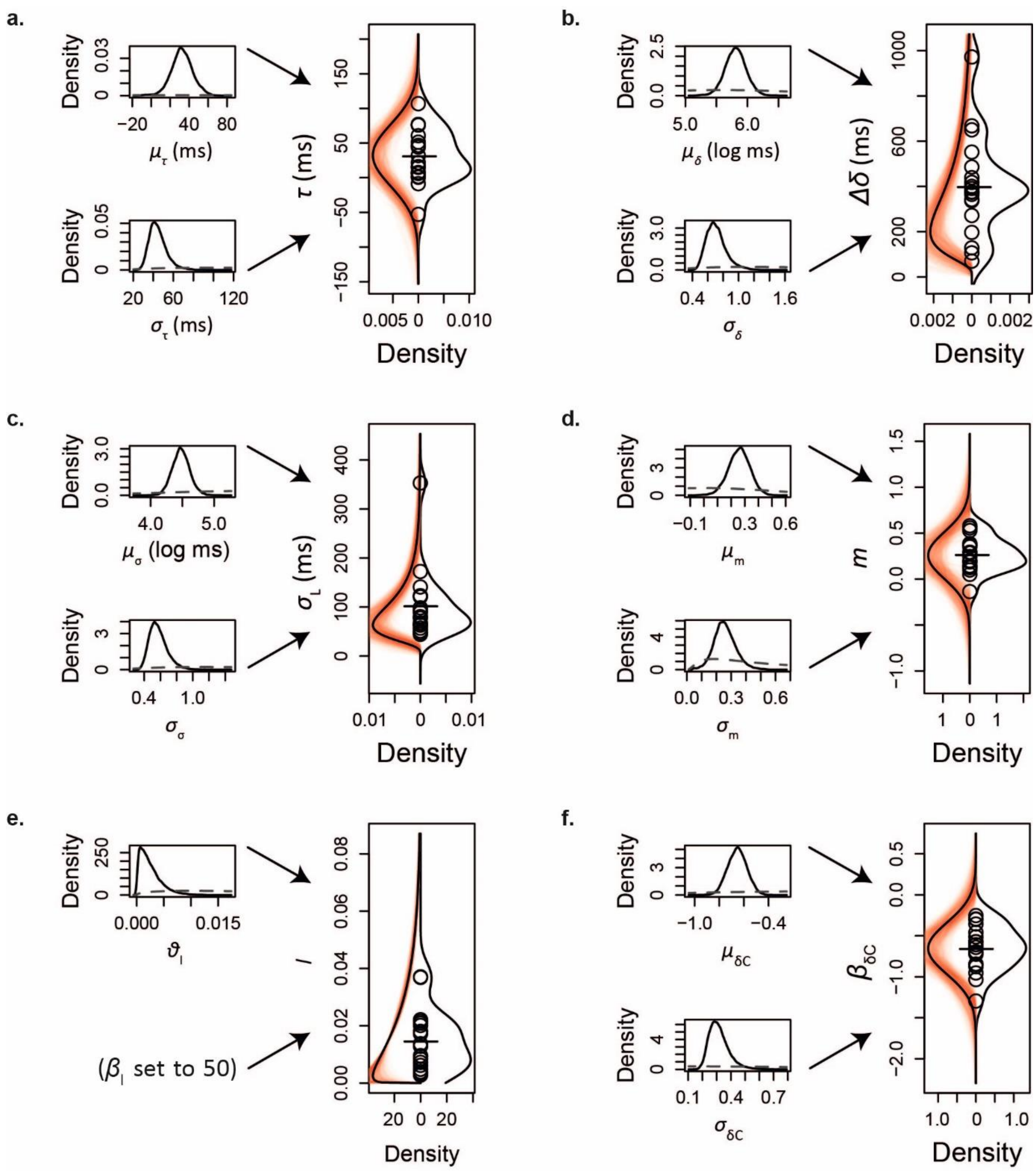

Figure C1. Summary of selected AT-A-GLANCE (criterial-variant) model parameter estimates. In each panel, smaller graphs on the left illustrate hyperpriors (dashed grey) and kernel-density estimates of posteriors (solid black) for group-level parameters. Values from these posteriors parameterise distributions predicting variation in participant-level parameters across the population (right-hand hourglass plots, left lobes; black line derived from mean of posteriors, red shading derived from entire posterior to illustrate uncertainty). Within these hourglass plots, individual estimates of *participant-level parameters are shown as black circles, their mean as a solid horizontal line, and a kernel-density estimate based on these estimates completes the hourglass plot as the right-hand lobe.* ***(a)*** *Group-level parameters μ_{τ} and σ_{τ} which describe the (normal) distribution of the participant-level parameter τ. This in turn describes the central tendency of the simultaneity function.* ***(b)*** *Group-level parameters μ_{δ} and σ_{δ} which describe the (lognormal) distribution of the participant-level parameter $\Delta\delta$. This in turn describes the width of the simultaneity function.* ***(c)*** *Group-level parameters μ_{σ} and σ_{σ} which describe the (lognormal) distribution of the participant-level parameter σ_{L}. This in turn describes the (inverse) slope of the simultaneity function's left flank.* ***(d)*** *Group-level parameters μ_{m} and σ_{m} which describe the (normal) distribution of the participant-level parameter m. The exponent of m is multiplied by σ_{L} in order to yield the (inverse) slope of the simultaneity function's right flank. Hence the group-mean value illustrated here implies σ_{H} was typically around 1.3 times as large as σ_{L}.* ***(e)*** *Group-level parameter ϑ_{l} which fixes the mode of the (beta) distribution of the participant-level parameter l. This in turn describes the (half) lapse rate defining lower/upper bounds on the simultaneity function.* ***(f)*** *Group-level parameters $\mu_{\delta\mathrm{C}}$ and $\sigma_{\delta\mathrm{C}}$ which describe the (normal) distribution of the participant-level coefficient $\beta_{\delta C}$. This is in turn exponentiated to form a multiplier quantifying how $\Delta\delta$ changes in the "be conservative" condition of the experiment (see main text Results section for further details relating to interpreting this coefficient).*

Appendix D: Parameter recovery simulations

To check that our methods were capable of adequately recovering model parameters, we simulated our experiment. We drew binomial-distributed random responses based on model-predicted “proportion judged synchronous” values for 19 participants in three conditions each with 9 SOAs and 30 trials per SOA. This was done based on known parameter values for each of our three (criterial-variant) models. We drew these parameter values at random from distributions that approximated those we had estimated for the population when fitting the models to our actual data. For example, when assessing parameter recovery for the AT-A-GLANCE model, the τ parameter for each simulated participant was drawn from a distribution similar to that shown as black against a red background in the hourglass plot of Appendix C Figure C1a, and so on for other parameters. Simulated data were then fit using the model that had generated them via a slightly truncated version of the same procedure that we applied to real data for our main analyses (with 5000 rather than 10000 post warmup samples per chain, to reduce computation time).

Figures D1 to D3 show actual vs. recovered parameter values (alongside the ideal lines of equality) for the criterial AT-A-GLANCE, ELA, and MCD models respectively. Parameters are in general recovered fairly well based on the numbers of trials and fitting procedures used in our experiment.

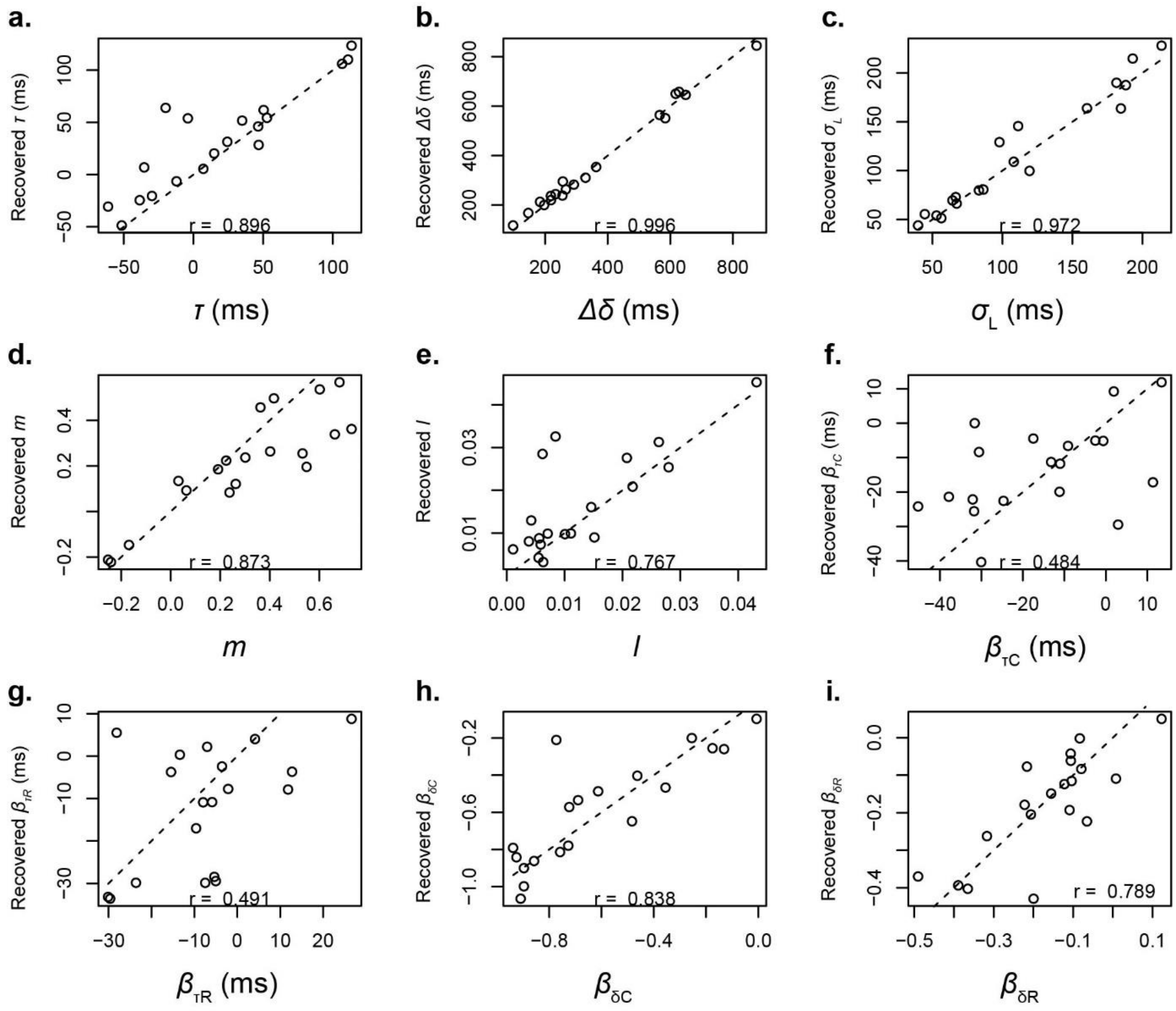

Figure D1. Parameter recovery simulation with criterial AT-A-GLANCE as the generating model. Dashed black line indicates equality for generative and recovered parameters; r = Pearson correlation coefficient. ***(a-e)*** *Model parameters describing baseline performance. These affect the psychometric function's midpoint, width, left-hand (inverse) slope, change in right-hand relative to left-hand (inverse) slope, and (half) lapse rate, respectively.* ***(f-i)*** *β Model parameters describing changes in position (τ) and width ($\Delta\delta$) of the psychometric function in the Conservative and Rebound conditions.*

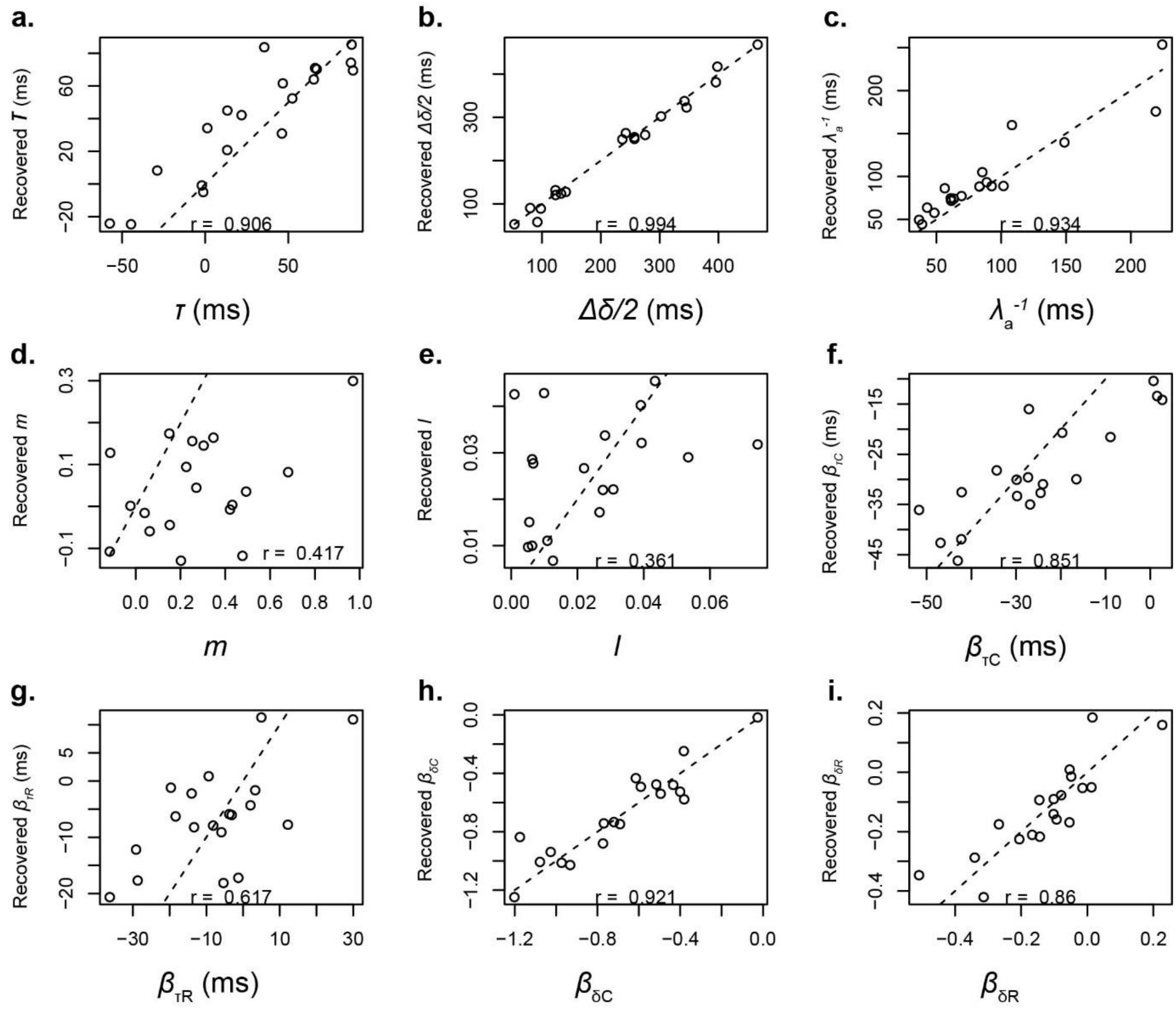

Figure D2. Parameter recovery simulation with criterial ELA as the generative model. Dashed black line indicates equality for generative and recovered parameters; r = Pearson correlation coefficient. ***(a-e)*** *Model parameters describing baseline performance. These affect the psychometric function's midpoint (τ), width ($\Delta\delta$), shape (λ_a^{-1} and m), and (half) lapse rate (l).* ***(f-i)*** *β Model parameters describing changes in position (τ) and width ($\Delta\delta$) of the psychometric function in the Conservative and Rebound conditions.*

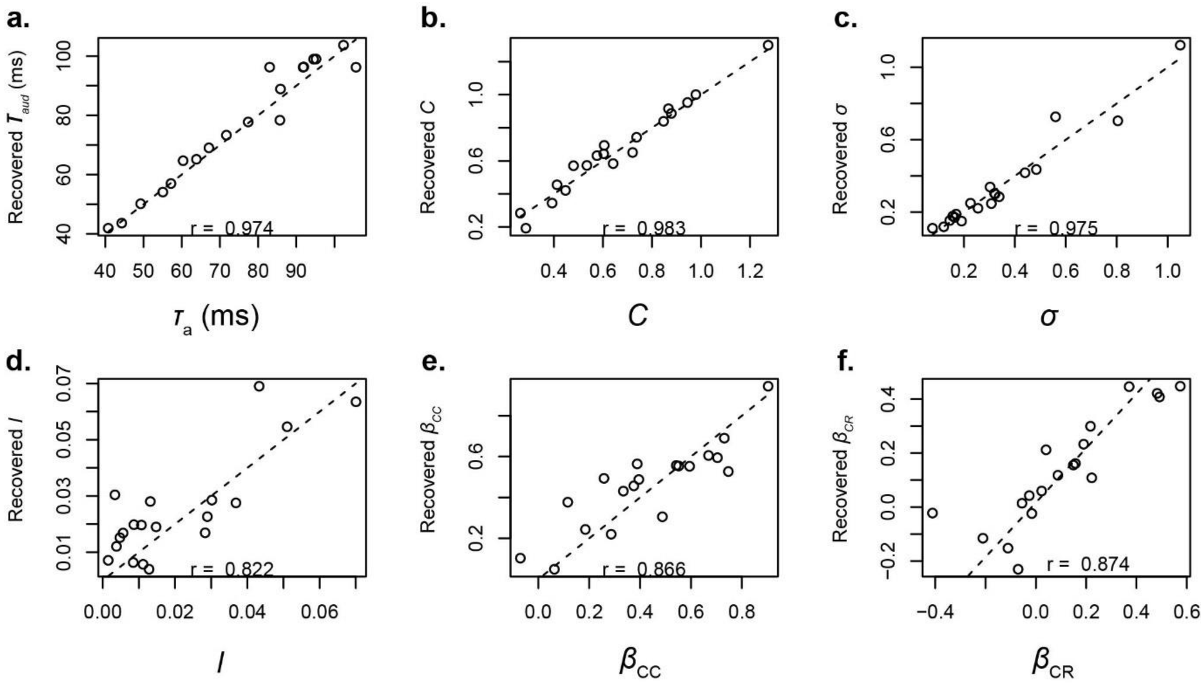

Figure D3. Parameter recovery simulation with criterial MCD as the generative model. Dashed black line indicates equality for generative and recovered parameters; r = Pearson correlation coefficient. ***(a-d)*** *Model parameters describing baseline performance. These affect the psychometric function via the model's auditory-filter time constant (τ_a), decision criterion (C), noise (σ), and (half) lapse rate (l).* ***(e-f)*** *β Model parameters describing changes in the decision criterion (C) affecting the psychometric function in the Conservative and Rebound conditions respectively.*

Appendix E: Ability of PSIS-LOO metric to compensate model complexity and discriminate true from false models

Appendix D, above, describes how we used each of our three (criterial-variant) models to create a simulated data set and fit that data set with the generative (i.e. true) model in order to assess parameter recovery. Further to this, we additionally recorded PSIS-LOO as a measure of goodness of fit (as per our main data analysis, but without additional leave-one-out substitution for Pareto ks > 1.0 to reduce computation time; hence a somewhat noisier approach to goodness-of-fit estimation). We then fit both of the alternative (i.e. false) models to that same simulated data and recorded PSIS-LOO for them in the same way. Finally, we repeated the whole procedure for a second run.

The resulting PSIS-LOO values are shown in Table E1. AT-A-GLANCE and ELA have identical numbers of free parameters. AT-A-GLANCE yields higher values of PSIS-LOO compared to ELA when it is the generative model (as expected). PSIS-LOO is more similar between these models when ELA is generative, although ELA wins (significantly) on one of the two runs. These results suggest that AT-A-GLANCE may be better able to mimic ELA than vice versa, at least with our procedures. The MCD model has less free parameters than both AT-A-GLANCE and ELA. As PSIS-LOO is intended to estimate goodness of fit while taking appropriate account of model complexity, MCD should nonetheless outperform the other two models when it is generative. It indeed scores significantly better, suggesting that the PSIS-LOO metric is working as intended in the current context and favouring a parametrically simpler generative model over more complex (but false) alternatives.

Table E1. Comparison of PSIS-LOO values between generative and non-generative models (two simulated experiments per model). Standard errors are shown in brackets. The asterisk (*) denotes a significant difference (z test p<.05) between a false model and the generative model for that simulated data set.

Data-generating model	Model fitted to data		
	AT-A-GLANCE	ELA	MCD
AT-A-GLANCE	-1059.0 (23.6)	-1110.1 (24.6)*	-1175.9 (31.4)*
	-1003.3 (22.2)	-1050.6 (24.6)*	-1200.0 (38.2)*
ELA	-1070.4 (20.0)	-1072.9 (19.9)	-1160.7 (26.8)*
	-1078.6 (19.7)*	-1061.8 (18.9)	-1150.9 (24.9)*
MCD	-1015.2 (23.1)*	-1024.0 (23.6)*	-968.5 (22.1)
	-1053.1 (23.0)*	-1086.6 (24.4)*	-992.8 (18.9)