| 1        | The best fitting of three contemporary observer models  |
|----------|---|
| 2        | reveals how participants' strategy influences the window of   |
| 3        | subjective synchrony  |
| 4        |   |
| 5        | Abbreviated Title: Modelling the window of subjective synchrony   |
| 7        | Kielan Yarrow <sup>1*</sup> , Joshua A. Solomon <sup>2</sup> , Derek H. Arnold <sup>3</sup> & Warrick Roseboom <sup>4,5</sup> |
| 8        |   |
| 9        | <sup>1</sup> Department of Psychology, City, University of London, London EC1V OHB  |
| 10       | <sup>2</sup> Centre for applied vision research, City, University of London, London EC1V OHB                                  |
| 11       | <sup>3</sup> School of Psychology, The University of Queensland, Brisbane QLD 4072  |
| 12       | <sup>4</sup> Department of Informatics. School of Engineering and Informatics. University of Sussex. Falmer.                  |
| 13       | Brighton BN1 901  |
| 11       | <sup>5</sup> School of Bsychology, University of Sussay, Falmer, Prighton, PN1 00H  |
| 14<br>15 | School of Psychology, Oniversity of Sussex, Funner, Brighton, BN1 9QH   |
| 16<br>17 | * Author for correspondence:  |
| 18       | Kielan Yarrow,  |
| 19       | Rhind Building,   |
| 20       | City, University of London,   |
| 21<br>22 | Northampton Square,   |
| 23       |   |
| 24       | Tel: +44 (0)20 7040 8530  |
| 25<br>26 | Email: <u>kielan.yarrow.1@city.ac.uk</u>  |
| 27       | Word count (main text excluding abstract, references & appendices): 12,212  |

#### 29 Abstract

30 When experimenters vary the timing between two intersensory events, and participants judge their 31 simultaneity, an inverse-U-shaped psychometric function is obtained. Typically, this *simultaneity* 32 *function* is first fitted with a model for each participant separately, before best-fitting parameters 33 are utilised (for example compared across conditions) in the second stage of a two-step inferential 34 procedure. Often, simultaneity-function width is interpreted as representing sensitivity to 35 asynchrony, and/or ascribed theoretical equivalence to a window of multisensory temporal binding. 36 Here, we instead fit a single (principled) multilevel model to data from the entire group and across 37 several conditions at once. By asking 20 participants to sometimes be more conservative in their 38 judgments, we demonstrate how the width of the simultaneity function is prone to strategic change 39 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 40 41 subjective asynchrony, and a third deriving this decision from the correlation between filtered 42 responses to sensory inputs) we find that the first model, which hypothesises, in particular, Gaussian 43 latency noise and difficulty maintaining the stability of decision criteria across trials, is most plausible 44 for these data. 45

#### 46 Keywords

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

48

#### 49 Public Significance

50 Psychologists have made their competing theories about how humans are able to perceive the 51 relative timing of events concrete by formulating mathematical models that attempt to describe 52 behaviour in specific experimental tasks. Here, we focus on one such task and show that people's 53 reports about simultaneity are inherently subjective, as implied by several current models. We also 54 find that the best-performing of these models explains inconsistencies when responding repeatedly 55 to objectively identical pairs of events by positing inconsistencies in both the time it takes for neural 56 messages to propagate through the brain, and how those messages are then interpreted to form a 57 decision.

59

60

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

61 The case has been made that the late eighteenth-century study of individual differences in 62 the time at which two events appear simultaneous was actually the founding question for 63 experimental psychology (Mollon & Perkins, 1996). Interest in this topic endures, and various tasks 64 have been developed over the years to help probe the human sense of relative time. In one such 65 task, known as the simultaneity judgment (SJ), participants are exposed to pairs of stimuli separated 66 by a range of asynchronies, and must judge each such pair to be either simultaneous or not (e.g. 67 Allan, 1975a). In the intermodal variant of this task, the two stimuli affect different senses, most 68 typically vision and audition.

69 This intermodal simultaneity-judgment task has proved popular with researchers for at least 70 three reasons. Firstly, for those whose primary interest is in understanding the mechanisms by which 71 we perceive relative time, the intersensory task seems to require the use of a specifically temporal 72 system, rather than allowing participants to fall back on alternative intramodal cues that are 73 processed by specialist systems, such as visual motion detectors (Cass & Van der Burg, 2014). 74 Secondly, participants seem to find the simultaneity-judgment task less onerous to perform than the 75 most popular alternative, the temporal order judgment (TOJ) task (Love et al., 2013). Thirdly, 76 temporal coincidence provides a powerful cue that events originating from different sensory 77 modalities have a common cause. Hence the determination of synchrony seems a necessary step 78 towards achieving another important cognitive operation: Multisensory integration. Indeed, such 79 integration is often found across a limited range of sub-second physical asynchronies, supporting the 80 concept of a temporal binding window within which multisensory integration can occur (Diederich & 81 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.

91 In this paper, we have two broad aims. The first is to make an initial determination regarding 92 which current model of the simultaneity judgment shows most promise. This necessitates that we 93 review several models. In so doing, we also provide groundwork for our second goal, which is to 94 caution researchers against making uncritical interpretations regarding summary measures, 95 particularly relating to the width of the simultaneity function. With these goals in mind, the 96 remainder of the introduction will progress as follows. First, we outline recent practice with regard 97 to the analysis of simultaneity judgments and highlight some interpretative issues. Next, we describe 98 three models of the simultaneity judgment (García-Pérez & Alcalá-Quintana, 2012a; Parise & Ernst, 99 2016; Yarrow et al., 2011). We then conclude the introduction by outlining an experiment that 100 provides a suitable data set with which to both compare models and demonstrate the dependence 101 of the simultaneity function on strategic decisions made by the participant.

102

#### 103 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.

111 Common practice is to fit the judgments from each participant / condition with a single such 112 sigmoidal psychometric function. The parameters of this function will then have meaning in relation 113 to the underlying model that justifies their use – for example, the mean of a fitted cumulative 114 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. 115 Alternatively, all participants and conditions can be fitted at the same time within a multilevel model 116 (Goldstein & McDonald, 1988). Such models acknowledge the clustering of individual data points 117 118 (here, responses within participants) and explicitly model random variation across clusters (here, 119 differences between participants across the group; Moscatelli et al., 2012; Prins & Kingdom, 2018).

120 In the case of the simultaneity judgment, properly formulated models of the psychometric 121 function (e.g. Schneider & Bavelier, 2003) seem not to have been widely appreciated. Principled 122 models do exist for simultaneity judgments, and relevant authors have sometimes made model-123 fitting code available, at least for fits to a single participant/condition at a time (Alcalá-Quintana & 124 García-Pérez, 2013; Yarrow et al., 2016; Yarrow, 2018). However, a tradition has emerged in which 125 researchers (including ourselves) instead resort to fitting a descriptive function that has no basis as a 126 model of participants' actual behaviour (for example, Roseboom & Arnold, 2011).<sup>1</sup>

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

<sup>&</sup>lt;sup>1</sup> 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.

138 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 139 140 conditions as indicative of differences in the temporal sensitivity of integration processes (e.g. Chen 141 et al., 2017; Foucher et al., 2007; Habets et al., 2017; Hillock et al., 2011; Lee & Noppeney, 2011; 142 Marsicano et al., 2022; Navarra & Fernández-Prieto, 2020; Noel et al., 2017; Scarpina et al., 2016; 143 Stevenson et al., 2014; Zampini et al., 2005). While of considerable interest, we believe that much of 144 this work does not include sufficiently explicit caveats about the processes that might generate the 145 window of simultaneity, potentially misrepresenting the underlying cause(s) of differences between 146 conditions/groups. It seems to us that this summary measure has poor face validity to measure the 147 conceptually distinct temporal-binding window. Hence one of our goals here is to advocate more 148 explicit recognition of the fundamentally subjective nature of the window derived from simultaneity 149 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.

164

#### 165 **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

179 shown for the example of a -50 ms SOA. Participants judge the trial as synchronous when the 180 subjective SOA falls between two decision criteria (solid red region). For the AT-A-GLANCE model 181 only, variable shading around the criteria indicates additional criterion noise; each criterion is most 182 likely to be placed where the shading is darkest, but varies across trials. (c) The MCD model has 183 sequential filtering operations on sensory inputs which lead to a signal that represents the temporal 184 cross-correlation between inputs (MCD<sub>corr</sub>). This signal is assumed to accrue Gaussian noise, and a 185 single criterion is applied, such that trials yielding values of (noisy) MCD<sub>Corr</sub> above this criterion are 186 judged simultaneous. (Note that the x-axis of the 3D inset differs from parts a and b – the 187 relationship between objective SOA and MCD<sub>corr</sub>, which is not shown, is roughly inverse-U shaped) (df) Solid black lines show resulting simultaneity functions. In each case, the point calculated in parts a-188 189 c is highlighted. Other points on the function are obtained in the same way. Dashed black lines show 190 what happens if parameters describing decision criteria are changed to model more conservative 191 behaviour. Dotted grey lines show predictions if criteria are assumed to reflect a hard threshold for 192 the perception of synchrony, so cannot be changed, but participants still attempt to reduce their use 193 of the synchronous response. (g) Replot of data from parts d-f illustrating how a hard-threshold 194 account predicts a linear relationship between proportion judged simultaneous in Baseline and 195 Conservative conditions (regardless of further modelling assumptions) whereas models in which 196 decision criteria change generally predict non-linearity. See main text for further details.

197

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.

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

212 The first model, which we term "AT-A-GLANCE" (Approximation To A Gaussian Latency And 213 Noisy Criteria Equation; Yarrow et al., 2011) assumes that latency noise – trial-by-trial changes in 214 the time taken for the neural responses to propagate through the brain to the decision centre – is 215 Gaussian in shape. On its own, this form of noise generates a symmetric simultaneity function. 216 However, it is further assumed that decision criteria are not held perfectly stable, but rather vary 217 from trial to trial (Ulrich, 1987), introducing a further source of noise that can differ for the two sides 218 of the psychometric function. If criterion noise is greater for one side of the psychometric function 219 (for example when discriminating simultaneous from sound-lags-light stimuli) than for the other (for 220 example discriminating simultaneous from sound-leads-light stimuli) the slope of the function will be 221 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
 <a href="https://kielanyarrow.github.io/MyPage/Code.html">https://kielanyarrow.github.io/MyPage/Code.html</a> (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).

#### 237 Exponential Latency Alone model of simultaneity

238 Our second model also hails from the independent-channels family (García-Pérez & Alcalá-239 Quintana, 2012a). We term it "ELA" (Exponential Latency Alone). Rather than assuming Gaussian 240 latency noise, this model assumes that each signal's propagation times through the brain can be 241 better described using an exponential distribution. A judgment is again formed at the hypothetical 242 decision centre by placing bounding criteria on the resulting distribution of subjective differences in 243 arrival times. However, unlike AT-A-GLANCE, these criteria are stable across trials. If each signal gives 244 rise to a different exponential distribution of arrival times (for example the distribution is tighter for 245 auditory than visual signals) this leads to asymmetry in the resulting psychometric function. Leaving 246 aside lapses, this model also uses four parameters per participant: A rate parameter for each 247 exponential distribution, which affect the slopes of the simultaneity function, and two parameters 248 determining its position and width. When each participant completes both simultaneity and 249 temporal order judgment tasks, a simultaneous fit of this model to all tasks at once has been shown 250 to provide a viable account of behaviour (García-Pérez & Alcalá-Quintana, 2012a; García-Pérez & 251 Alcalá-Quintana, 2015). The model has also been used to show how the inclusion of lapse and keying 252 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).

## 255 Multisensory Correlation Detector model applied to simultaneity judgments

256 The final model we implement here has a different background. This MCD (multisensory 257 correlation detector) model (Parise & Ernst, 2016) is broadly analogous to popular accounts of 258 motion detection in vision (Fujisaki & Nishida, 2007). It builds on earlier ideas that perceived 259 simultaneity might be a function of the degree of overlap between the internal responses to two 260 stimuli, which can be thought of as temporally low-pass filtered versions of the input (Burr et al., 261 2009; Stelmach & Herdman, 1991). In the MCD model, each signal first passes through a modality-262 specific filter. The output of one modality is then multiplied by an additionally filtered version of the 263 other, and vice versa. Finally, the two resultant signals are multiplied together and then integrated 264 over the interval of time immediately following presentation of the stimuli in order to provide a 265 single quantity (MCD<sub>Corr</sub>) that represents perceived synchrony. To yield a categorical response, this 266 quantity is compared to a single criterion, above which synchrony is reported. Noise for this 267 judgment accrues from Gaussian variation in either the strength of MCD<sub>Corr</sub> (which is otherwise 268 deterministic) or the placement of the criterion across trials (these two ideas yield identical 269 predictions so cannot be discriminated).

270 Leaving aside lapses, this model has five parameters (three filter time constants, a criterion, 271 and a noise term). However, it has traditionally been fitted to simultaneity judgments by fixing the 272 filter time constants based on additional data sets and utilising a two-parameter generalised linear 273 model. Based on our explorations regarding the recoverability of model parameters, we opted to 274 build upon a three-core-parameter (plus lapses) fit. We fixed both the second-stage filter time 275 constant and the visual-filter time constant, but allowed the auditory-filter time constant to vary. 276 Changing the ratio of time constants for the two unisensory filters generates asymmetry in the 277 psychometric function (and also a correlated shift in its central tendency) while the noise term

| 278 | affects slopes, and the criterion term affects width. The model can be explored via our               |
|-----|---|
| 279 | aforementioned Shiny App. Example applications of this model include explaining data from a range     |
| 280 | of synchrony tasks with stimuli that employ both simple and complex temporal profiles (for example    |
| 281 | simultaneity judgments, temporal-order judgments, and various judgments about correlated and          |
| 282 | uncorrelated trains of stimuli). The model has also helped provide viable neural loci for the process |
| 283 | of cross-correlating multisensory stimuli (Pesnot Lerousseau et al., 2022) and, with slight           |
| 284 | modification, helped explain the effect of visual luminance on simultaneity judgments and temporal-   |
| 285 | order judgments (Horsfall et al., 2021).  |

286

# 287 Testing whether strategy influences the simultaneity function

288 Having summarised the candidate models, we can now move on to introduce an 289 experimental manipulation. In previous sections we have alluded to the idea that categorical reports 290 (for example "simultaneous") might be generated by applying decision criteria to underlying 291 perceptual representations that are continuous. The underlying representation could be an arrival-292 time difference (as assumed in the AT-A-GLANCE and ELA models) or a cross-correlation of filtered 293 inputs (as per the MCD model). Conscious experience could reflect these continuous quantities, but 294 making binary decisions would require that the underlying representation is categorised using some 295 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.

304 It is straightforward to test whether the criteria applied to the simultaneity-judgment task 305 when participants first walk into the laboratory are hard thresholds of this kind. We can do it, for 306 example, by introducing a condition in which participants are asked to reduce their use of the 307 simultaneous response option (Yarrow, 2018). The models that we have described include 308 parameters which could be allowed to change in such a condition in order to represent a change of 309 decision criteria. This is illustrated by the dashed black lines in Figure 1 parts d-f.

310 But how would we know that decision criteria had really changed, rather than merely 311 seeming to change as an artefact of fitting an inappropriate model to data? The answer involves 312 predicting what would happen if thresholds obtained at baseline remained hard. With a "be 313 conservative" instruction encouraging a limited number of synchronous responses over the many 314 trials of the experiment, observers would sometimes need to report asynchrony despite perceiving 315 synchrony. The result would be a proportional reduction of the predicted psychometric function 316 (Figure 1 d-f, dotted grey lines). It is straightforward to embed such an account in an observer 317 model, as an alternative parameter that can change in conservative conditions instead of decision 318 criteria.

319 One concern with such an approach would be that it involves comparing two variants of an 320 observer model, and such models are mere approximations of reality. For this reason, we 321 additionally consider a test of the hard-threshold account that does not depend on any particular 322 observer model. To this end, we can reframe how we visualise the data. Rather than plotting 323 proportion judged synchronous in both baseline and conservative conditions against SOA (as per 324 Figure 1 panels d-f) we can consider proportion judged synchronous in the conservative condition as 325 a function of proportion judged synchronous in the baseline condition (Figure 1g). If the thresholds 326 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.

337 First, we will test for an anticipated violation of linearity in the function predicting 338 proportion judged synchronous in the conservative condition from proportion judged synchronous 339 in the baseline condition. Next, for each of the three simultaneity-judgment models we have 340 outlined, we will fit two multilevel model variants to data from all participants and all conditions at 341 once. In the first, parameters relating to decision criteria will be allowed to vary across conditions. In 342 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 343 344 parameters) for the former model variants compared to the latter, which would further support the 345 idea that simultaneity judgments are in part strategic. We will also compare goodness of fit across 346 our three types of simultaneity-judgment model (AT-A-GLANCE, ELA, and MCD). Few if any 347 comparisons of this type exist, so we are interested to see which of these models provides a 348 prediction that is closest to our data, and thus receives greatest support.

350

# Method

# 351 Participants

| 352 | This study comprises a secondary analysis of data published previously as a pre-print                       |
|-----|---|
| 353 | (Yarrow & Roseboom, 2017). An opportunity sample of twenty observers, all naïve to the                      |
| 354 | experimental purpose, participated in early 2017. Written informed consent was acquired from all            |
| 355 | participants prior to the experiment, which was approved by the University of Sussex ethics                 |
| 356 | committee. Participants received £5 per hour or course credit as compensation for their time.               |
| 357 | Demographics were not retained with the dataset, but the sample was recruited from the same                 |
| 358 | predominantly undergraduate student panel, at around the same time, as that reported in                     |
| 359 | Roseboom (2019), which contained 60% females with a mean age around 22 (SD 5) years.                        |
| 360 | The current work addresses both the originally intended research question (the effect of                    |
| 361 | strategy on simultaneity judgments), but via a more comprehensive analysis, and an additional               |
| 362 | research question (by comparing different models of the simultaneity judgment). To our knowledge,           |
| 363 | the most relevant prior observation regarding the effect of strategy came from a single-case study          |
| 364 | (subsequently described in Yarrow, 2018). This indicated an effect that was large in absolute terms         |
| 365 | but, with N = 1, could not be normed to a standardised measure of effect size. Hence the sample size        |
| 366 | was selected (prior to the initiation of data collection) based on prevailing norms for simultaneity-       |
| 367 | judgment studies with similar designs. Data from one participant were not included in the final             |
| 368 | analysis (see data analysis, below). For a paired-samples t-test, the remaining N = 19 participants         |
| 369 | yield a-priori power of 91% to detect a large (Cohen's $d = 0.8$ ) effect size (at two-tailed alpha = .05). |
| 370 | With regard to our second research question, relating to model comparison, we provide data                  |
| 371 | relevant to evaluating power in Appendix E, where 5/6 simulations (using our sample size) yielded a         |
| 372 | significant difference between the generative model and each of the non-generative models.                  |

#### 374 Apparatus and stimuli

375 Participants sat in a quiet, bright room. Visual stimuli were displayed on either an liyama 376 Vision Master Pro 203 or LaCie Electron 22 Blue II monitor, both with a resolution of 1024 x 768 377 pixels and refresh rate of 100 Hz. The monitor was positioned at a viewing distance of approximately 378 57 cm. Audio signals were presented binaurally through Sennheiser HDA 280 PRO headphones. 379 Stimulus generation and presentation was controlled through Psychtoolbox 3 (Brainard, 1997) run in 380 MatLab (Mathworks, USA) on a desktop PC. Participants responded using the computer keyboard. 381 Visual events were luminance-modulated Gaussian blobs ( $\sigma = 1.5$  degrees of visual angle 382 (dva)) displayed against a grey background (approximately 38 cd/m<sup>2</sup>). Peak blob luminance was 383 approximately 76 cd/m<sup>2</sup>. A fixation square (white, approximately 76 cd/m<sup>2</sup>, subtending 0.25 dva) 384 was presented centrally. The Gaussian blob was centred 3 dva above the fixation square. The visual 385 stimulus was presented for one frame approximating 10 ms in duration. Auditory signals were a 10 386 ms amplitude pulse of 1500 Hz sine-wave carrier at approximately 55 db SPL.

387

#### 388 Design and procedures

389 The experiment consisted of six sessions. Each took approximately seven minutes to

390 complete. In each session, participants were presented with a sequence of 135 audio-visual

391 presentations. Each presentation consisted of visual and auditory events presented with one of nine

392 pseudo-randomly interleaved stimulus-onset asynchronies (SOAs;  $\Delta t \in$ 

 $\{-400 \text{ ms}, -200 \text{ ms}, -100 \text{ ms}, -50 \text{ ms}, 0, 50 \text{ ms}, 100 \text{ ms}, 200 \text{ ms}, 400 \text{ ms}\}, \text{ where positive values}$ 

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

396 response as to whether the auditory and visual events had occurred at the same time

397 (synchronously; up cursor key) or not (asynchronously; down cursor key).

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.

For the first two experimental sessions (270 trials), these were the only instructions given.

403

398

404 Data analysis

#### 405 Modelling approach and software

406 We opted to apply Bayesian multilevel models, which we consider the most principled way 407 to treat these data and test our hypotheses. In recent years, multilevel models have seen 408 widespread advocacy and adoption across diverse fields including neuroscience (Aarts et al., 2014) 409 and psychology (Barr et al., 2013). This includes the active promotion of their use to analyse data 410 from psychophysical tasks (e.g. Moscatelli et al., 2012). For standard (sigmoidal) psychometric 411 functions, packages such as the Palamedes toolbox (Prins & Kingdom, 2018) offer multilevel 412 approaches "off the shelf". However, we are not aware of any such option for those interested in 413 modelling simultaneity judgments. We therefore fit Bayesian multilevel models using the Stan 414 programming language interfaced from R (R Core Team, 2021) via the RStan package (Stan 415 Development Team 2020; 2022). We share our commented code (see Transparency and Openness 416 subsection, below) as a potential template for other researchers interested in developing bespoke 417 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 418 419 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 420 421 our reported model fits use 1000 warmup iterations followed by 10,000 post-warmup iterations per 422 chain.

#### 423 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.

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

432 participants

433 In our model-based analyses, we utilised three classes of multilevel simultaneity-judgment 434 model, each with two variants: A strategic variant which allows one or more parameters that 435 represent participants' decision criteria to change across experimental conditions, and a hard-436 threshold variant which instead allows the psychometric function to show proportional reduction. 437 This proportional reduction mimics an attempt to reduce use of the simultaneous response option 438 when all stimuli judged simultaneous give rise to the exact same perceptual experience, as the only 439 option for the participant would then be to reply "asynchronous" at random to some stimuli they 440 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  $\beta$ . 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  $\beta$  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  $\beta$  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 ( $\mu_{\beta}$ ) coefficients. We provide statements of significance similar to frequentist null-hypothesis testing based on whether the 95% credible interval contains 0 or 1.

452 We also incorporated *posterior predictive checks* (Lambert, 2018). The posterior predictive 453 distribution of any one of our  $\beta$  coefficients tells us what we can expect for future participants, and in combination with its standard error (which equals its  $SD/\sqrt{N}$ ) it can provide an alternative means 454 455 of evaluating differences from 0 or 1, via a single-sample t-test.<sup>2</sup> We also used a posterior predictive 456 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 457 lower than that for a random binomial draw conditioned on model parameters. If the model is 458 459 correct for an individual, this Bayesian P value should be around .5, with higher values indicating 460 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. 461 462 Finally, we wished to compare the two model variants (for each class of simultaneity-463 judgment model) to one another in order to evaluate which of our hypotheses received greater 464 support. We can estimate a model's out-of-sample goodness of fit via leave-one-out cross validation, 465 but this is very time consuming. Hence, we instead used an estimate of leave-one-out cross 466 validation via Pareto smoothed importance sampling (Vehtari et al., 2017), known as PSIS-LOO. This 467 measure is based on the log-likelihood of the model given the data, but utilises the full posterior 468 distribution of parameter values in estimating goodness of fit, and corrects for the number of model 469 parameters in a more nuanced fashion than better-known metrics such as the Akaike and deviance

470 information criteria (AIC, DIC).

<sup>&</sup>lt;sup>2</sup> 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.

471 PSIS-LOO was estimated and compared between model variants (and indeed between 472 classes of model) using functions from the R package, LOO, and z tests (which are based on the 473 difference between models in units of the standard error of this difference). Although PSIS-LOO can 474 be multiplied by -2 to give an AIC-like value where low is best, we don't bother to apply this 475 transform, so report negative values, where higher is better. PSIS-LOO for the whole model is found 476 by summing log likelihood estimates for each data point. The LOO package provides diagnostics and 477 outputs which together indicate the number and positions of data points for which the PSIS-LOO 478 estimate is potentially inaccurate. We therefore replaced a small number of data points considered 479 "very bad" (Pareto k value > 1.0) via direct leave-one-out cross validation, and also report the 480 number of estimates considered "bad" (Pareto k > 0.7), which we elected not to replace due to the 481 heavy computational demands of doing so.

#### 482 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
- 486 <u>https://city.figshare.com/articles/software/Code\_and\_data\_accompanying\_The\_best\_fitting\_of\_thr</u>
- 487 <u>ee\_contemporary\_observer\_models\_reveals\_how\_participants\_strategy\_influences\_the\_window\_o</u>
- 488 <u>f\_subjective\_synchrony\_/20495652</u>.

490

#### Results

#### 22

491

## 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.

503

504 From Figure 2 panel a, it appears that participants reduced their use of the synchronous 505 response option when asked to "be conservative", but not in a manner that was proportional across 506 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 514 therefore sought evidence to reject this null hypothesis by fitting a multilevel model with a zero 515 516 intercept, but fixed and random effects for both slope and, critically, the exponent of the power 517 function. This specification allows variation in both slope and exponent for each participant.<sup>3</sup> It yielded an estimated group-mean exponent of 1.58 (solid black line in Figure 2b) with a credible 518 519 interval (1.33-1.85) that did not include 1. This result provides grounds to reject the hard-threshold 520 account. We next moved to more fully characterise our data via three observer models of 521 simultaneity-judgment behaviour, starting with AT-A-GLANCE.

# 522 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

<sup>&</sup>lt;sup>3</sup> We used a binomial data model, so very slightly corrected the prediction (to be, at the individual level, y = 0.00001+0.99999\*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).

529 distribution (with 30 trials) for the number of "simultaneous" responses ( $S_{X\Delta t}$ ) from each participant 530 in each condition (X = B, X = C, and X = R, for baseline, conservative, and rebound conditions, 531 respectively) with each objective SOA ( $\Delta t$ ):

532 (1) 
$$S_{X\Delta t} \sim B(30, l + p_{X\Delta t} - lp_{X\Delta t})$$
,

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

535 (2) 
$$p_{X\Delta t} = \Phi\left[\frac{\Delta t - \tau + \beta_{\tau X} D_X - \exp(\beta_{\delta X} D_X) \Delta \delta/2}{\sigma_{\rm L}}\right] - \Phi\left[\frac{\Delta t - \tau + \beta_{\tau X} D_X + \exp(\beta_{\delta X} D_X) \Delta \delta/2}{\exp(m)\sigma_{\rm L}}\right]$$

In Equation 2 exp is the exponential function,  $\Phi$  is the standard normal cumulative distribution function, and  $D_X$  is a "dummy" or indicator variable that equals 1 if and only if X = C or X = R. The remaining 8 symbols ( $\tau$ ,  $\Delta\delta$ ,  $\beta_{\tau C}$ ,  $\beta_{\tau R}$ ,  $\beta_{\delta C}$ ,  $\beta_{\delta R}$ ,  $\sigma_L$ , 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.

541 In this model, the  $\tau$  and  $\Delta\delta$  parameters capture the midpoint and width (respectively) of each participant's psychometric function. They provide an alternative (and mathematically 542 543 equivalent) way of describing the positions of two decision criteria (because  $\Delta\delta$  is the distance 544 between these criteria, which are centred on  $\tau$ ). Hence, our hypothesis that decision criteria vary 545 with task instructions can be tested by allowing these two parameters to vary across conditions. For this purpose, four parameters,  $\beta_{\tau C}$ ,  $\beta_{\tau R}$ ,  $\beta_{\delta C}$  and  $\beta_{\delta R}$ , represent changes between conditions 546 547 (compared to baseline) with the first subscript representing the parameter being adjusted and the second representing the <u>C</u>onservative and <u>R</u>ebound conditions. The  $\sigma_{\rm L}$  and m parameters describe 548 549 noise affecting the left flank of the psychometric function, and the noisiness of the right flank 550 relative to the left flank (m of 0 indicating an identical magnitude of noise), respectively. Like the 551 lapse-rate parameter l, these final two parameters were assumed constant across experimental conditions. 552

553 For our second "hard-threshold" AT-A-GLANCE model variant, the four parameters 554 permitting changes across conditions were replaced with just two ( $\beta_{\rm C}$  and  $\beta_{\rm R}$ ), each describing a 555 proportional reduction in the number of trials judged synchronous for a given condition.

556 For both variants, our multilevel models additionally estimated random variation across the 557 group via group-level distributions from which the individual-level parameters were drawn. This 558 required a further 17 (or 13) parameters (for criterial and hard-threshold variants, respectively). For 559 example, we estimated, for the Gaussian group-level distribution of individual  $\tau$  parameters, a group 560 mean ( $\mu_{\tau}$ ) and standard deviation ( $\sigma_{\tau}$ ). Similarly, for the group-level distribution of *changes* in  $\tau$  from 561 the baseline to the conservative condition, we estimated a further group mean ( $\mu_{\tau C}$ ) and standard 562 deviation ( $\sigma_{\tau C}$ ). Full details are provided in Appendix A (with group-level distributions visualised in 563 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

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.

586 Table 1. Summary of models.

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

587

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 leaveone-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.<sup>4</sup>

597 We can therefore reasonably compare PSIS-LOO values between the two model variants 598 that formalise different theories regarding how participants respond to instructions across our three 599 experimental conditions. The difference in PSIS-LOO of 57.9, with a standard error of 19.9, implies 600 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 601 602 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 603 604 the synchronous response.

605 Some evidence that criterial AT-A-GLANCE is in fact a plausible account of these data (in an 606 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 607 data, so here, higher than a binomial distribution would be expected to yield).<sup>5</sup> As indicated in Table 608 609 1, for the criterial model, only one out of 19 participants had a Bayesian P value above .95, which is 610 around the chance expectation if the model is correct. However, for the hard-threshold model, 6 participants showed overdispersion of this magnitude. 611 In Figure 4, we plot results from a subset of four participants – those showing the lowest and 612 613 highest overdispersion, so effectively the best and worst fits, for each of the two different variants of

614 th

highest overdispersion, so effectively the best and worst fits, for each of the two different variants o the AT-A-GLANCE multilevel model. The hard-threshold model cannot capture a common pattern in

<sup>&</sup>lt;sup>4</sup> 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.

<sup>&</sup>lt;sup>5</sup> 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.





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.

625

626 We can also consider exactly how the parameters of the significantly more successful 627 criterial variant of the AT-A-GLANCE model have changed across the three experimental conditions. 628 Two parameters were allowed to change. The first,  $\tau$ , describes the point midway between decision 629 criteria, and is comparable with the commonly reported "point of subjective simultaneity". In the baseline condition, the mean of its group-level distribution ( $\mu_{\tau}$ ) was 32 ms (95% credible interval 10 630 to 55). This implies a group-average bias to report simultaneity more when sound lags light than vice 631 632 versa (individual values for all participants can be seen in Appendix C Figure C1a). However, this bias 633 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 634 635 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 636 637 response strategy. Importantly, we also observed that the mean of the distribution describing 638 changes in psychometric function width ( $\mu_{\delta C}$ ) was -0.65 (credible interval -0.80 to -0.50; see also 639 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, 640 641 exponentiation provides more meaningful units: The distance between decision criteria has changed 642 (shrunk) by an average factor of 0.52. This means that participants are making simultaneous responses for a reduced range of audio-visual timings.<sup>6</sup> 643 644 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

<sup>&</sup>lt;sup>6</sup> 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/v19.

646 simultaneous. Both have moved inwards in the conservative condition, but this change is less 647 pronounced for the low criterion. It showed an average shift of +70 ms, but with a credible interval 648 from -11 ms to 326 ms that hence includes zero. The high criterion showed an average shift of -115 649 ms (credible interval -373 ms to -33 ms). Regardless of how the criteria have been parameterized, 650 their shifts suggest that participants appropriately adjusted their decision-making strategies in 651 accordance with the instructions to be more conservative. More specifically, participants made more 652 of an adjustment regarding how light-leading stimuli should be classified compared to how sound-653 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.<sup>7</sup> 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.

660 The ELA and MCD models

661 In addition to the above-described results for the AT-A-GLANCE model, we tested two 662 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 663 664 overlap in neural responses, rather than arrival times at a neurocognitive hub. Mathematical details 665 appear in Appendix A. Returning to Table 1, it is apparent that the AT-A-GLANCE criterial-model 666 variant shows substantially better goodness-of-fit metrics compared to all other models. A statistical 667 comparison suggests that these differences are meaningful. We focus on the generally better-668 performing criterial variants of each class of model. A difference in PSIS-LOO of 44.7 (with a standard

<sup>&</sup>lt;sup>7</sup> 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 669 criterial ELA model (frequentist two-tailed z test, z = 3.06, p = .002). The difference was even more 670 striking for criterial AT-A-GLANCE versus criterial MCD (PSIS-LOO difference = 85.5, SE 20.5, z = 4.17, 671 672 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).<sup>8</sup> As a methodological check (and test of model 673 674 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 675 676 (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 677 678 was not. The MCD model was beaten convincingly by both AT-A-GLANCE and ELA when they were 679 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 680 intended in the current context. 681 682 Because both ELA and MCD provided significantly less compelling descriptions of the data 683 relative to AT-A-GLANCE, we will spend less time describing their detailed results. However, Figure 5 684 provides some insights into why these models performed less well. The figure plots fits from all

685 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).

<sup>&</sup>lt;sup>8</sup> 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

692 represent 1000 samples from the full posterior (each plotted with high transparency) in order to

693 illustrate uncertainty in the model prediction, and error bars (which represent 95% binomial
694 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 soundleading 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).

700 For the ELA class of model, like AT-A-GLANCE, the criterial variant showed better goodness 701 of fit compared to the hard-threshold variant. However, the difference in PSIS-LOO (39.6, with a 702 standard error of 26.9) was not statistically compelling (z = 1.47, frequentist two-tailed p = .141). For 703 the MCD model, the trend actually reversed (with the hard-threshold variant outperforming the 704 criterial variant) although the magnitude of the difference was very small in relation to estimation 705 error (PSIS-LOO difference 4.2, SE 25.3, z = 0.17, p = .868). However, because AT-A-GLANCE provided 706 a significantly better overall account of the data, and agreed with our non-model-based test (see 707 section Non model-based assessment of the hard-threshold account), we give priority to that model 708 when interpreting differences between model variants in relation to our experimental hypothesis. 709

710

#### Discussion

711 In this paper, we have presented data from an experiment requiring judgments about the 712 simultaneity of audio-visual pairs. Participants made these judgments under conditions that either 713 let them freely decide how to behave, or encouraged them to be conservative in their use of the 714 simultaneous response option. Data were then fitted with two variants of each of three multilevel 715 observer models of simultaneity judgments. The two variants of each model represented different 716 hypotheses about how participants would attempt to address the experimental instruction. If 717 flexible criteria exist and determine which subjective stimulus patterns are classified as 718 simultaneous, participants would be expected to adjust those criteria when asked to be 719 conservative. If no such criteria were being applied in the first place, consistent with truly binary 720 perceptual experiences arising from some hard neurocognitive thresholding mechanism, participants 721 would have two choices. They might either fail to adjust their behaviour at all, or sometimes respond 722 "non-simultaneous" even to perceptually compelling experiences of simultaneity in order to meet 723 experimental demands.

724 Our first observation is that of the three classes of model that we tested, AT-A-GLANCE 725 (Approximation to a Gaussian Latency Independent Noisy Criteria Equation; Yarrow et al., 2011), a 726 variant of the general-threshold family of models (Ulrich, 1987) provided the best account of the 727 data. This is, to our knowledge, the first time a direct comparison between two or more of these 728 models has been attempted. Given AT-A-GLANCE's success (in both relative and absolute terms) we 729 prioritised this model for the evaluation of our experimental hypothesis regarding the existence of 730 decision criteria. Of AT-A-GLANCE's two variants, the criterial variant, corresponding with the 731 hypothesis that participants were applying flexible internal decision criteria in order to categorise 732 stimuli as simultaneous or not, significantly outperformed the hard-threshold variant. This was in 733 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 734

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).

740

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

742 examination

743 The most successful of our models, AT-A-GLANCE, has much in common with the second

744 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

r50 simultaneity criteria are viewed as being constant or variable from trial to trial.<sup>9</sup>

751 AT-A-GLANCE's particular combination (Gaussian latency noise and variable criteria) was

- 752 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

<sup>&</sup>lt;sup>9</sup> 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  $\delta$  (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.
755 (and gives rise to both a computationally efficient model prediction and a posterior likelihood 756 surface that is highly amenable to search and characterisation). Meanwhile, AT-A-GLANCE's use of 757 Gaussian noise must be strictly incorrect to the extent that it permits propagation times to be 758 negative. Something in between the two (for example some shifted gamma distribution aside from 759 the exponential), probably with additional criterion noise, therefore holds conceptual appeal. It 760 would be plausible when considering the nature of neuronal transmission, and offer the possibility 761 of separately characterising the noise associated with each stimulus. However, there are practical 762 issues to consider that make this avenue of research challenging. Model parameters would likely 763 become more degenerate (meaning it would be more difficult to recover a unique value for each). 764 There would also be increasingly subtle differences between the psychometric functions that 765 different blended models would predict.

766

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

769 The multisensory correlation detector (MCD) model (Parise & Ernst, 2016) is a highly 770 attractive one. It offers both a lower level of abstraction relative to both AT-A-GLANCE and ELA, and 771 the promise of immediate application to a wider range of experimental tasks, such as those involving 772 complex trains of stimuli. However, MCD was markedly less successful in describing our data set. 773 This might in part be because it does not offer independent mechanisms to affect the central 774 tendency of the simultaneity function and the relative slopes of its two flanks. However, the more 775 fundamental problem seems to have been that under the multisensory correlation detector model 776 the derived decision variable (MCD<sub>Corr</sub>) effectively throws away information about the sign of the 777 SOA. Hence any change in the (single) decision criterion that is applied to this signal has similar 778 effects at both sides of the simultaneity function. In contrast to this, some participants seem to 779 selectively adjust decisions more for sound-lagging compared to sound-leading stimuli.

180 It is difficult to envisage a simple change that might resolve this problem, because it arises 1781 from a core feature of the MCD architecture. Hence, at this point we conclude that an *MCD*<sub>corr</sub>-like 1782 signal cannot be the only source of information determining how participants judge simultaneity in 1783 the simultaneity-judgment task (although it might contribute). In saying this we do however 1784 acknowledge that there could be systemic differences between how timing decisions are made 1785 between different individuals or groups. A focus on group-level summary measures comparing 1786 distinct models might obscure any such differences.

787

#### 788 Interpretations based on simultaneity judgments should bear in mind the task's criterion-

789 dependent nature

790 Broadly, there are two mechanisms which might be envisaged as a limit on an observer's 791 precision (or on their sensitivity or acuity, which are synonymous terms). The first is internal noise. 792 The second is an inflexible (hard) thresholding mechanism which irretrievably reduces a continuous 793 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 794 795 supported model (AT-A-GLANCE) provide converging evidence regarding whether a hard threshold 796 should be inferred from SJ data. Both favour the alternative idea that judgments of simultaneity are 797 formed by classifying a continuous underlying signal according to decisional criteria. The fact that 798 these decisional criteria reverted only partially in the rebound condition suggests that, for many 799 participants, at least three criterial settings were attainable. It might also imply that the settings 800 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).

811 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 812 813 the existence of such a structural threshold within (or above, in the case of the multisensory 814 correlation detector model) the criterial range that is naturally obtained. The current methodology 815 might be extended to address this kind of question, or at least to place a limit on the magnitude of 816 any structural threshold, by forcing ever-more conservative behaviour through stricter rationing of 817 the simultaneous response option. Ideally, this would be done with highly motivated participants 818 and closely spaced SOAs. Such an approach would complement previous attempts to test hard-819 threshold accounts for relative time. For example, Baron (1971) offered a first and second guess 820 about which of two synchronous and one preceding stimulus came first, and assessed the degree to 821 which second guesses (following an initial failure) yielded above-chance performance. That 822 approach, which focussed specifically on triads of intramodal (visual) stimuli, ruled out certain kinds 823 of hard-threshold account (Allan, 1975b). These include accounts in which noise in performance 824 comes relative to a background sampling process (for example the moving moment model of Stroud, 825 1956). However, it also provided evidence against independent-channels models without any 826 thresholds. With the addition of appropriate model comparison, it might be used to formally assess 827 remaining alternatives, such as models with hard ("low") thresholds accompanied by sensory noise 828 (Swets et al., 1961). Our results here indicate that if, in the audio-visual case, a hard threshold does 829 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 thiscombination of instruction and task does not reveal what that threshold might be.

832 These findings regarding the important role that decision criteria play stand in contrast to 833 the widespread interpretation of simultaneity-function width as an unambiguous measure of the 834 precision of multisensory integration (e.g. Chen et al., 2017; Foucher et al., 2007; Habets et al., 2017; 835 Hillock et al., 2011; Lee & Noppeney, 2011; Marsicano et al., 2022; Navarra & Fernández-Prieto, 836 2020; Noel et al., 2017; Scarpina et al., 2016; Stevenson et al., 2014; Zampini et al., 2005). We have 837 already indicated how our results show that simultaneity-function width in uninstructed baseline 838 conditions is not a measure of a hard sensory threshold, if indeed one exists. That leaves the 839 question of whether it is a measure of internal noise. It is plausible, and even predicted by some 840 accounts of what an optimal observer is trying to do, that there might be a correlation between the 841 spacing of decision criteria and the noise underlying perception. Sensitivity should often inform 842 strategy, potentially linking these conceptually distinct measures (Magnotti et al., 2013). However, 843 researchers should be mindful that any difference between the widths of simultaneity functions 844 would then only be indirectly driven by differences in, for example, the consistency of arrival times 845 at a central comparator. We note that the naïve expectation that wider windows of perceived 846 simultaneity should predict less or worse multisensory integration has received somewhat mixed 847 support (for example Stevenson et al., 2018). Viewing the width of the simultaneity function from 848 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 855 necessarily reflect perceptual limitations. Moreover, given that there are easily derived model 856 parameters that have a better claim to represent internal noise in multisensory perception (for 857 example those affecting the slope of the simultaneity function, such as  $\sigma$  parameters for AT-A-858 GLANCE and MCD, and  $\lambda$  parameters for ELA) we suggest that these measures should more often 859 take the limelight.

860 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 861 862 measures the researcher's definition of integration/binding, rather than the participant's definition 863 of simultaneity. For example, consider the redundant-signals effect. This is a reaction-time 864 advantage obtained over and above a statistical facilitation when responding to audio-visual pairs 865 rather than their individual components (e.g. Colonius & Diederich, 2004; Diederich & Colonius, 866 2015; Hershenson, 1962; Miller, 1982; Raab, 1962; Schwarz, 2006). It is measurable when the audio-867 visual pair is near synchronous. The redundant-signals effect implies multisensory integration has 868 occurred: The two signals have interacted in a way that modifies behaviour relative to the sum of 869 their individual effects. Furthermore, the timing between component signals is an important 870 determinant. It would, in our opinion, be a reasonable task with which to quantify the dependency 871 of multisensory integration upon the timing between signals. By contrast, at least at face value, the 872 range of audio-visual timing relationships over which I declare two signals to be simultaneous has 873 little claim to measure the range of values at which my brain integrates/binds them in order to 874 generate a multisensory advantage. We would argue that the near-ubiquitous (but extremely 875 leading) term "temporal binding window" should be replaced with something more neutral, like 876 "window of subjective simultaneity" when summarising the results of simultaneity-judgment 877 studies.

#### 879 Bayesian multilevel modelling is a complex but powerful approach to analysing simultaneity

#### 880 judgment experiments

881 We could have fitted our models using the common two-step approach of first fitting a 882 model to each individual, and then assessing group differences using a procedure such as the *t*-test. 883 Multilevel models have advantages over such a two-stage analysis. Perhaps most importantly, by 884 fitting all participants at once, multilevel models can generate "shrinkage", whereby well-estimated 885 participants help constrain parameter estimates for less well-estimated participants (Lambert, 886 2018). The result can be more powerful, robust and reliable estimation that generally performs 887 better in out-of-sample prediction (Aarts et al., 2014; Lambert, 2018; Moscatelli et al., 2012). 888 Shrinkage may also have practical value in a field where it is common to reject participants on the 889 basis that their data are inadequate to generate reliable parameter estimates (and in which pre-890 registration of exclusion criteria is not yet the norm). If there are ways to reduce the number of 891 participants who have to be excluded, we should probably adopt them.

892 Bayesian models additionally encourage the explicit specification of sensible priors, or rather 893 hyper-priors in the case of multilevel models. When used judiciously, these should further enhance 894 the reliability of recovered parameters. They also make use of the full distribution of plausible 895 parameter values from the posterior when assessing the goodness of a model's fit, rather than 896 relying exclusively on the mode of the posterior, as per maximum likelihood estimation. Compared 897 to popular metrics like the Akaike information criterion (AIC), Bayesian metrics (for example 898 estimation of leave-one-out cross validation via Pareto smoothed importance sampling; Vehtari et 899 al., 2017) are likely to provide a better estimate of a model's out-of-sample predictive accuracy, and 900 thus a fairer means of comparing models with different architectures (Lambert, 2018). Here, we 901 have demonstrated how such Bayesian multilevel modelling can be used to evaluate whether model 902 parameters change across conditions, and to test more complex hypotheses via the instantiation of 903 these hypotheses as competing models.

904 We hope that the code accompanying this paper, in concert with Appendix A, can act as a 905 template for other researchers interested in using similar approaches. Although we have focussed 906 on the popular simultaneity-judgment task, there is a range of tasks that generate non-sigmoidal 907 psychometric functions that might benefit from bespoke multilevel modelling along these lines. In 908 the realm of time perception, these include judgments about which of two intervals contained a 909 more synchronous signal (Yarrow et al., 2016) or whether the duration of a test stimulus matched 910 that of a pre-learnt standard, often referred to as temporal generalization (Bausenhart et al., 2018; 911 García-Pérez, 2014). There are also analogous tasks in other fields (e.g. García-Pérez & Peli, 2014; 912 Morgan et al., 2013). Nonetheless, we must acknowledge that because of the need for bespoke 913 coding, the time investment for this type of analysis exceeds that associated with the application of 914 simpler tests (such as t-tests) as a second-stage inferential step. For example, we have only 915 illustrated a test of whether/how parameters change across a single experimental factor, via dummy 916 coding. Implementing factorial designs would require technical knowledge regarding how to 917 implement the equivalent of ANOVA models within a multilevel model framework, for example the 918 proper use of effects coding. However, we doubt this is beyond the abilities of the average 919 quantitively minded researcher.

920 We are additionally mindful that the benefits of shrinkage that accrue from the multilevel 921 approach are premised on the correctness of modelling assumptions regarding group-level 922 distributions. For example, in the AT-A-GLANCE and ELA models we assumed a normal distribution 923 for the group when modelling the  $\tau$  parameter. This describes the central tendency of individual 924 simultaneity-judgment functions, so is the parameter most conceptually akin to the commonly 925 reported "point of subjective simultaneity". But what if the population actually consists of a number 926 of distinct sub-groups, perhaps reflecting very different task strategies or neurological types? Then, 927 the implied uniformity, in terms of the computational processes underlying timing decisions, would 928 be incorrect, and shrinkage toward the group mean could be inappropriate. This would be most 929 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.

935

936 Further caveats, limitations, and constraints on generality

937 There are several reasons to be cautious regarding our conclusions here, which are derived 938 from work with a necessarily limited scope. Firstly, our study lacked a fundamental feature of well-939 designed repeated-measures experiments – the counterbalancing of the order of experimental 940 conditions to remove practice and fatigue effects. This was justified by our desire to capture 941 instinctive behaviour in the simultaneity-judgment task before meddling with people's strategies, 942 but it implies that differences between conditions might be contaminated by learning effects. We 943 acknowledge this problem, but note that the inclusion of the rebound condition provides some 944 reassurance that the main driver of differences between conditions was the instruction we provided. 945 Secondly, with the exception of our non-model-based test of the hard-threshold account,

946 our conclusions follow from the exact choices we made when implementing simultaneity-judgment 947 models, and strictly cannot be generalised beyond that context. For example, we used a single lapse-948 rate parameter l, but might reasonably have used two such parameters to capture a bias towards 949 one or other response when guessing, as has been implemented by the authors of ELA (García-Pérez 950 & Alcalá-Quintana, 2012b). We gave all three models identical flexibility in this regard, but it is 951 possible that their relative statuses would have changed had we made different choices. The same 952 follows for other decisions, including our choice of hyperpriors (but see Appendix C) and the 953 parameters that were allowed to change across conditions. Allowing only criteria to vary was largely 954 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.

959 In terms of scope, we have tested only a limited range of models, and used only the audio-960 visual simultaneity-judgment task with austere stimuli. As noted above, a variety of blended or 961 modified models could be entertained. Furthermore, there is at least one recently advocated class of 962 model relevant to simultaneity judgments that we have ignored: Population-code (sometimes called 963 labelled-line) models (Roach et al., 2011; Roseboom et al., 2015; Yarrow et al., 2015). However, 964 there were reasons for leaving this class of model out. In the absence of some manipulation based 965 on sensory adaptation, its basic simultaneity-judgment prediction is very similar to AT-A-GLANCE, 966 but without the noisy criteria aspect. However, to deal with established differences in slope for the 967 two sides of the simultaneity function, one would need to add something like noisy criteria. This 968 remains entirely within the spirit of a population-code model, as the population of neurones simply 969 supplies an estimate of the represented quantity, in this case subjective SOA, and is agnostic with 970 regard to further steps to formulate a binary decision. Indeed, population-code accounts of the 971 simultaneity judgment are perhaps best viewed as a more fleshed-out representational stage within 972 an independent channels / general-threshold framework (Yarrow & Arnold, 2016). To this extent, 973 the current result can be viewed as supportive of a population code (plus noisy criteria) as much as 974 of AT-A-GLANCE.

975 We have also focussed here exclusively on modelling the simultaneity-judgment task. Of 976 course, more general models are typically preferable to models which explain only one particular 977 phenomenon. It is possible to extend models like those we test here to simultaneously account for 978 data from multiple tasks. One example is Diederich and Colonius' (2015) simultaneous account of 979 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.

## 989 Conclusion

990 Here, we have demonstrated how to investigate experimental questions addressed using the 991 simultaneity-judgment task by fitting Bayesian multilevel models, illustrating this approach with 992 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 993 994 model akin to AT-A-GLANCE, because the ultimate arbitrator between theories should probably be 995 how well they predict out of sample data, and AT-A-GLANCE performed best in this regard. We have 996 also shown that performance on the simultaneity-judgment task reflects an interpretation by the 997 participant based on malleable decision criteria. It is these criteria that determine the width of the 998 simultaneity function, and hence the window of subjective simultaneity. Thus, because of its 999 strategic nature, this window casts only a thin light on multisensory temporal integration/binding 1000 processes, and should be interpreted with caution. Although no universal remedy, changes in 1001 measures that directly assess internal noise seem more pertinent when drawing conclusions about 1002 the causes underlying perceptual differences between clinical and other groups.

1004

## References

| 1005 | Aarts, E., Verhage, M., Veenvliet, J. V., Dolan, C. V., & Van Der Sluis, S. (2014). A solution to           |
|------|---|
| 1006 | dependency: using multilevel analysis to accommodate nested data. Nature Neuroscience,                      |
| 1007 | <i>17</i> (4), 491-496.   |
| 1008 | Alcalá-Quintana, R., & García-Pérez, M. A. (2013). Fitting model-based psychometric functions to            |
| 1009 | simultaneity and temporal-order judgment data: MATLAB and R routines. Behavior Research                     |
| 1010 | Methods, 45(4), 972-998.  |
| 1011 | Allan, L. G. (1975a). The relationship between judgments of successiveness and judgments of order.          |
| 1012 | Perception and Psychophysics, 18(1), 29-36.   |
| 1013 | Allan, L. G. (1975b). Second guesses and the attention-switching model for successiveness                   |
| 1014 | discrimination. Perception & Psychophysics, 17, 65–68.  |
| 1015 | Baron, J. (1971). The threshold for successiveness. <i>Perception &amp; Psychophysics, 10</i> (4), 201-207. |
| 1016 | Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory       |
| 1017 | hypothesis testing: Keep it maximal. Journal of Memory and Language, 68(3), 255-278.                        |
| 1018 | Bausenhart, K. M., Di Luca, M., & Ulrich, R. (2018). Assessing duration discrimination: Psychophysical      |
| 1019 | methods and psychometric function analysis. Timing and Time Perception: Procedures,                         |
| 1020 | Measures, & Applications (pp. 52-78). Brill.  |
| 1021 | Bonnet, E., Masson, G. S., & Desantis, A. (2022). What over When in causal agency: Causal                   |
| 1022 | experience prioritizes outcome prediction over temporal priority. Consciousness and Cognition,              |
| 1023 | <i>104</i> , 103378. https://doi.org/10.1016/j.concog.2022.103378   |
| 1024 | 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.
- 1029 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.
- 1032 Colonius, H., & Diederich, A. (2004). Multisensory interaction in saccadic reaction time: a time-

1033 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.
- 1036 Foucher, J. R., Lacambre, M., Pham, B., Giersch, A., & Elliott, M. A. (2007). Low time resolution in
- 1037 schizophrenia: lengthened windows of simultaneity for visual, auditory and bimodal stimuli.
- 1038 Schizophrenia Research, 97(1), 118-127.
- 1039 Freeman, E. D., Ipser, A., Palmbaha, A., Paunoiu, D., Brown, P., Lambert, C., Leff, A., & Driver, J.

1040 (2013). Sight and sound out of synch: Fragmentation and renormalisation of audiovisual

- 1041 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.
- 1044 García-Pérez, M. A., & Alcalá-Quintana, R. (2012a). On the discrepant results in synchrony judgment
- 1045 and temporal-order judgment tasks: a quantitative model. Psychonomic Bulletin & Review,

1046 *19*(5), 820-846.

- 1047 García-Pérez, M. A., & Alcalá-Quintana, R. (2012b). Response errors explain the failure of
- 1048 independent-channels models of perception of temporal order. *Frontiers in Psychology*, *3*, 94.
  1049 10.3389/fpsyg.2012.00094
- 1050 García-Pérez, M. A. (2014). Does time ever fly or slow down? The difficult interpretation of
- 1051 psychophysical data on time perception. *Frontiers in Human Neuroscience, 8*, 415.
- 1052 García-Pérez, M. A., & Alcalá-Quintana, R. (2015). Converging evidence that common timing
- 1053 processes underlie temporal-order and simultaneity judgments: A model-based analysis.
- 1054 Attention, Perception, & Psychophysics, 77(5), 1750-1766.
- 1055 García-Pérez, M. A., & Peli, E. (2014). The bisection point across variants of the task. *Attention,*
- 1056 *Perception, & Psychophysics, 76*(6), 1671-1697.
- 1057 Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*.
  1058 Cambridge university press.
- 1059 Goldstein, H., & McDonald, R. P. (1988). A general model for the analysis of multilevel data.
- 1060 *Psychometrika, 53*(4), 455-467.
- 1061 Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.
- 1062 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.
- 1066 Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: age-related
- 1067 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.
- 1070 Horsfall, R., Wuerger, S., & Meyer, G. (2021). Visual intensity-dependent response latencies predict
- 1071 perceived audio–visual simultaneity. *Journal of Mathematical Psychology, 100,* 102471.
- 1072 Johnston, P. R., Alain, C., & McIntosh, A. R. (2022). Individual Differences in Multisensory Processing
- 1073 Are Related to Broad Differences in the Balance of Local versus Distributed Information. *Journal*
- 1074 *of cognitive neuroscience*, 34(5), 846–863. https://doi.org/10.1162/jocn\_a\_01835
- 1075 Lambert, B. (2018). A student's guide to Bayesian statistics. Sage.
- 1076 Lee, H., & Noppeney, U. (2011). Long-term music training tunes how the brain temporally binds
- 1077 signals from multiple senses. *Proceedings of the National Academy of Sciences of the United*
- 1078 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.
- 1085 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.
- 1087 https://doi.org/10.1038/s41598-022-06503-1

| 1088 | Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in   |
|------|--|
| 1089 | superior colliculus neurons. I. Temporal factors. <i>Journal of Neuroscience, 7</i> (10), 3215-3229. |
| 1090 | Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. Cognitive    |
| 1091 | Psychology, 14(2), 247-279.  |
| 1092 | Mioli, A., Diolaiuti, F., Zangrandi, A., Orsini, P., Sebastiani, L., & Santarcangelo, E. L. (2021).  |
| 1093 | Multisensory Integration Is Modulated by Hypnotizability. The International journal of clinical      |
| 1094 | and experimental hypnosis, 69(2), 215–224. https://doi.org/10.1080/00207144.2021.1877089             |
| 1095 | Mollon, J. D., & Perkins, A. J. (1996). Errors of judgement at Greenwich in 1796. Nature, 380, 101-  |
| 1096 | 102.   |
| 1097 | Morgan, M. L. Melmoth, D., & Solomon, I. A. (2013). Linking hypotheses underlying Class A and Class  |

- 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.
- 1099 Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-

1100 level: the generalized linear mixed model. *Journal of Vision, 12*(11), 10.1167/12.11.26.

- 1101 10.1167/12.11.26 [doi]
- 1102 Navarra, J., & Fernández-Prieto, I. (2020). Perceptual association enhances intersensory temporal
  1103 precision. *Cognition, 194*, 104089.
- 1104 Noel, J., De Niear, M. A., Stevenson, R., Alais, D., & Wallace, M. T. (2017). Atypical rapid audio-visual
- 1105 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.

| 1108 | Pesnot Lerousseau, J., Parise, C. V., Ernst, M. O., & van Wassenhove, V. (2022). Multisensory        |  |  |  |
|------|--|--|--|--|
| 1109 | correlation computations in the human brain identified by a time-resolved encoding model.            |  |  |  |
| 1110 | Nature Communications, 13, 2489.   |  |  |  |
| 1111 | 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*,
  1113 1250.
- 1114 R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for
  1115 Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- 1116 Raab, D. H. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York*1117 *Academy of Sciences, 24*, 574-590.
- 1118 Roach, N. W., Heron, J., Whitaker, D., & McGraw, P. V. (2011). Asynchrony adaptation reveals neural
- 1119 population code for audio-visual timing. Proceedings of the Royal Society B: Biological Sciences,
- 1120 *278*(1710), 1314-1322. 10.1098/rspb.2010.1737
- 1121 Roseboom W. (2019). Serial dependence in timing perception. Journal of experimental psychology.
- 1122 Human perception and performance, 45(1), 100–110. https://doi.org/10.1037/xhp0000591
- 1123 Roseboom, W., & Arnold, D. H. (2011). Twice upon a time: multiple concurrent temporal
- recalibrations of audiovisual speech. *Psychological Science*, 22(7), 872-877.
- 1125 10.1177/0956797611413293
- Roseboom, W., Nishida, S., & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision, 9*(12), 4.
- 1128 Roseboom, W., Linares, D., & Nishida, S. (2015). Sensory adaptation for timing perception.
- 1129 Proceedings of the Royal Society of London B: Biological Sciences, 282(1805), 20142833.

| 1130 | Scarpina, F., Migliorati, D., Marzullo, P., Mauro, A., Scacchi, M., & Costantini, M. (2016). Altered |
|------|--|
| 1131 | multisensory temporal integration in obesity. Scientific Reports, 6, 28382. 10.1038/srep28382        |
| 1132 | [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.

1138 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

1141 Stelmach, L. B., & Herdman, C. M. (1991). Directed attention and perception of temporal order.

1142 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.

1146 Stevenson, R. A., Baum, S. H., Krueger, J., Newhouse, P. A., & Wallace, M. T. (2018). Links between

1147 temporal acuity and multisensory integration across life span. *Journal of Experimental* 

1148 *Psychology: Human Perception and Performance, 44*(1), 106.

1149 Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., & Wallace, M. T. (2014). The impact of

1150 multisensory integration deficits on speech perception in children with autism spectrum

disorders. *Frontiers in Psychology*, *5*, 379.

- 1152 Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., & Porter, N. R.
- 1153 (2002). When is Now? Perception of simultaneity. *Proceedings of the Royal Society of London*
- 1154Series 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.
- 1159 Ulrich, R. (1987). Threshold models of temporal-order judgments evaluated by a ternary response
- 1160 task. Perception and Psychophysics., 42(3), 224-239.
- 1161 van Eijk, R. L., Kohlrausch, A., Juola, J. F., & van de Par, S. (2008). Audiovisual synchrony and
- 1162 temporal order judgments: effects of experimental method and stimulus type. *Perception and*1163 *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.
- 1166 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).
- 1168 Bayesian Analysis, 16(2), 667-718.
- 1169 Venables, P. H. (1960). Periodicity in reaction time. *British Journal of Psychology*, *51*, 37-43.
- 1170 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.
- 1174 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.
- 1176 10.1016/j.concog.2011.07.003
- 1177 Yarrow, K. (2018). Collecting and interpreting judgments about perceived simultaneity: A model-
- 1178 fitting tutorial. In A. Vatakis, F. Balci, M. Di Luca, Á. Correa (Eds.), *Timing and Time Perception:*
- 1179 *Procedures, Measures, & Applications* (pp. 295-325). Brill.
- 1180 Yarrow, K., & Arnold, D. H. (2016). The Timing of Experiences: How Far Can We Get with Simple Brain
- 1181 Time Models? In B. Mölder, V. Arstila & P. Øhrstrøm (Eds.), *Philosophy and Psychology of Time*1182 (pp. 187-201). Springer.
- 1183 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.
- 1185 Yarrow, K., Martin, S. E., Di Costa, S., Solomon, J. A., & Arnold, D. H. (2016). A roving dual-
- 1186 presentation simultaneity-judgment task to estimate the point of subjective simultaneity.
- 1187 Frontiers in Psychology, 7, 416.
- 1188 Yarrow, K., Minaei, S., & Arnold, D. H. (2015). A model-based comparison of three theories of
- audiovisual temporal recalibration. *Cognitive Psychology*, *83*, 54-76.
- 1190 Yarrow, K., & Roseboom, W. (2017, October 31). Should multisensory temporal acuity be viewed
- 1191 through the window of perceived simultaneity? https://doi.org/10.31234/osf.io/zpkq7

- 1192 Yarrow, K., Sverdrup-Stueland, I., Roseboom, W., & Arnold, D. H. (2013). Sensorimotor Temporal
- 1193 Recalibration Within and Across Limbs. *Journal of Experimental Psychology: Human Perception*

1194 *and Performance, 39*(6), 1678-1689.

- 1195 Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments.
- 1196 *Perception & Psychophysics, 67*(3), 531-544.

1198

#### Appendix A: Multilevel model specifications

## 1199 AT-A-GLANCE model implementation

## 1200 Single-level AT-A-GLANCE

1201 Our first multilevel model built upon the AT-A-GLANCE four-parameter single-level observer 1202 model (Yarrow et al., 2011). Our description of that model here is more complete than in any of our 1203 previous papers and thus supersedes them. Under this account, the observer judges two stimuli 1204 simultaneous when the internal signals they generate arrive at a decision centre with a subjective 1205 SOA that is both above a (noisy) low criterion and below a (noisy) high criterion. Hence AT-A-GLANCE 1206 implies three normally distributed random variables: Two decision criteria ( $c_{\rm L}$  and  $c_{\rm H}$ ) used to 1207 demarcate successive judgments from simultaneous judgments, and the subjective SOA, s. These 1208 three random variables can be expressed as a single, trivariate normal random variable, with mean 1209 and variance:

1210 (A1) 
$$\boldsymbol{\mu} = \begin{pmatrix} \mu_{\rm L} \\ \mu_{\rm S} \\ \mu_{\rm H} \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{\rm L}^2 & \rho_{\rm LS}\sigma_{\rm L}\sigma_{\rm S} & \rho_{\rm LH}\sigma_{\rm L}\sigma_{\rm H} \\ \rho_{\rm LS}\sigma_{\rm L}\sigma_{\rm S} & \sigma_{\rm S}^2 & \rho_{\rm SH}\sigma_{\rm S}\sigma_{\rm H} \\ \rho_{\rm LH}\sigma_{\rm L}\sigma_{\rm H} & \rho_{\rm SH}\sigma_{\rm S}\sigma_{\rm H} & \sigma_{\rm H}^2 \end{pmatrix}$$

1211 Let  $f(c_L, s, c_H)$  denote its density. Then:

1212 (A2) 
$$P(C_{\rm L} < S < C_{\rm H}) = \int_{-\infty}^{\infty} dc_{\rm H} \int_{-\infty}^{C_{\rm H}} ds \int_{-\infty}^{S} dc_{\rm L} f(c_{\rm L}, s, c_{\rm H})$$

1213 Unfortunately, expressed in this way the (single-level) model has a hefty eight parameters 1214 (excluding  $\mu_{s}$ , which equals the experimental SOA). We can easily take a view regarding the  $\rho$ 1215 parameters, for example fix them to 0 for fully uncorrelated sources of noise, but Equation A1 is still 1216 slow to evaluate (we are not aware of a closed-form solution) and likely degenerate with regard to 1217 the three  $\sigma$  parameters (meaning they can trade off against each other to give near identical 1218 predictions). However, if we assume  $\rho_{LH} = 1$  and  $\rho_{LS} = \rho_{SH} = 0$ , implying correlated noise in the two 1219 criteria, a closed-form approximation is available:

1220 (A3) 
$$P(S|\Delta t) \approx \Phi\left(\frac{\Delta t - c_{\rm L}}{\sigma_{\rm L}}\right) - \Phi\left(\frac{\Delta t - c_{\rm H}}{\sigma_{\rm H}}\right)$$

1238

where S denotes the event that the observer responds "simultaneous",  $\Delta t$  is the SOA, and  $\Phi$ 1221 is the standard normal cumulative distribution function. The  $\sigma$  values quantify (inversely) the slope 1222 1223 on each side of the psychometric function. These are composite noise variables, used because they 1224 are formally identifiable in a model fit (meaning that they do not trade off perfectly) whereas the 1225 various psychological constructs that feed into them are not. Each  $\sigma$ , when squared, represents the sum of two sources of variance. The first, the variance of subjective SOAs ( $\sigma_s^2$  from Equation A1) is 1226 1227 itself derived from the (Gaussian) latency variance associated with each stimulus (if we assume 1228 uncorrelated sensory channels, it is their sum). This source contributes to the slope on both sides of 1229 the psychometric function (low and high). The second, the trial-by-trial (Gaussian) variance in a decision criterion ( $\sigma_{\rm L}^2$  or  $\sigma_{\rm H}^2$  from Equation A1) has a unique magnitude on each side of the function, 1230 1231 thus allowing the slopes to vary. Because Equation A3 is an approximation, we used a slower-to-1232 evaluate method when the approximation breaks down - this can be seen in our code as an "override" function.<sup>10</sup> 1233 1234 To this model, we first added a lapse parameter, l, such that participants are assumed to make 1235 an (unbiased) guess on a proportion of trials equalling 2l. This effectively forms upper and lower limits 1236 on the psychometric function at l and 1- l. We also opted to reparametrize  $c_{\rm L}$  and  $c_{\rm H}$  in terms of their midpoint and the distance between them.<sup>11</sup> We call these  $\tau$  and  $\Delta\delta$  respectively, to make the 1237 terminology comparable with near-equivalent parameters from a second observer model, ELA,

 $<sup>^{10}</sup>$  Equation A3 breaks down when the difference between  $\sigma_{
m L}$  and  $\sigma_{
m H}$  is large relative to the distance between  $c_{\rm H}$  and  $c_{\rm L}$ . 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.

<sup>&</sup>lt;sup>11</sup> 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.

1239 described later. We also reparametrized  $\sigma_{\rm H}$  by instead sampling the posterior based on *m*, the natural 1240 log of a quantity applied as a multiplier to  $\sigma_{\rm L}$  in order to determine  $\sigma_{\rm H}$ :

1241 (A4) 
$$\sigma_{\rm H} = \exp(m)\sigma_{\rm L}$$

1242 Hence the reparametrized single-level model incorporating lapsing becomes:

1243 (A5) 
$$P(S|\Delta t) \approx l + (1-l) \left[ \Phi\left(\frac{\Delta t - \tau + \Delta \delta/2}{\sigma_{\rm L}}\right) - \Phi\left(\frac{\Delta t - \tau - \Delta \delta/2}{\exp[m]\sigma_{\rm L}}\right) \right]$$

For completeness, we next implemented priors (and provide accompanying code) for a singlelevel 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.

## 1247 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 grouplevel 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).

1254 To the 95 participant-level parameters (coded as five vectors/arrays) we therefore added a 1255 set of group-level parameters. Multilevel models require that we specify a distribution (for example 1256 normal) which describes the way each participant-level parameter varies across the group. The 1257 parameters of these distributions are then estimated alongside the individual-level parameters: In 1258 effect, when determining the likelihood of a set of parameters for a particular participant, we consider 1259 both the likelihood of their data given their participant-level parameters, and the likelihood of those 1260 participant-level parameters given the group-level distribution from which they are presumed to be 1261 being drawn.

For the AT-A-GLANCE model, we specified a normal group-level distribution for parameter  $\tau$ , 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 ( $\mu_{\tau}$ ) and the standard deviation ( $\sigma_{\tau}$ ) of this distribution as parameters for estimation within the model.<sup>12</sup>

1268 For each group-level distribution parameter, Bayesian modelling encourages us to also 1269 specify a (hyper)prior distribution, based on our subject-specific knowledge. This is a somewhat 1270 uncomfortable step for those with a frequentist background, but hyperpriors can be made as 1271 uninformative/diffuse as the modeller desires (at least when considering just the untransformed 1272 parameter). Furthermore, the alternative perspective is quite hard to defend. It implies that any and 1273 all values for a group-level summary statistic such as the mean midpoint of perceptual synchrony are 1274 equally likely before we see our particular set of data. However, using  $\mu_{\tau}$  as an example, even in a 1275 case study of a patient with a specific relative-timing related pathological deficit, the reported point 1276 of subjective simultaneity was only +210 ms (Freeman et al., 2013). Hundreds of group averages of 1277 similar measures have been reported in the literature, and although we have not reviewed them all, 1278 we are confident that *all* are much closer to zero than to, say, ±1000 ms. 1279 Here, we utilised an (extremely diffuse) Cauchy hyperprior on  $\mu_{\tau}$ , with location of 0 and 1280 scale of 800 ms. Our code defaults to setting the former to the (unweighted) mean SOA in the data 1281 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

1282

hyperprior on  $\sigma_{\tau}$ , which should be zero-bounded, we used a lognormal distribution with  $\mu$  of 5.59

<sup>&</sup>lt;sup>12</sup> 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  $\tau$  and also the  $\beta$  coefficients representing changes across conditions (described later). Essentially, we modelled variation across the group using a standard normal distribution, then derived scaled values of  $\tau$  for each participant by multiplying this standardised variation by the group  $\sigma$  before adding the group  $\mu$ .

and  $\sigma$  of 1. The code defaults  $\mu$  to the natural log of one-third the range of asynchronies in the data, which, along with an  $\sigma$  of 1, for our data gives a right-skewed distribution with a mode of  $\approx$  100 ms. Note that the  $\mu$  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  $\tau$  to be normally distributed across the group, with a group mean vaguely near zero ms and a group SD vaguely near 100 ms.

1290 For  $\Delta\delta$ , the distance between the two judgment criteria, which is zero-bounded, we 1291 specified a lognormal group-level distribution and had the model estimate both parameters ( $\mu_{\delta}$  and 1292  $\sigma_{\delta}$ ). For hyperpriors on  $\mu_{\delta}$  and  $\sigma_{\delta}$ , we used normal and lognormal distributions respectively, the 1293 former with a  $\mu$  of 5.59 and  $\sigma$  of 1.4 and the latter with  $\mu$  of 1.1 and  $\sigma$  of 1 (the code again defaults to 1294 basing some of these on the range of asynchronies found in the data). This translates to expecting 1295  $\Delta\delta$  to vary across the group according to heavily right-skewed distribution with a mode vaguely near 1296 90 ms, but with hyperpriors giving plenty of scope for very different central tendencies and shapes 1297 to emerge.

1298 For  $\sigma_{\rm L}$ , the inverse slope of the left side of the psychometric function, we specified a 1299 lognormal group-level distribution (and hyperpriors on its two parameters,  $\mu_{\sigma}$  and  $\sigma_{\sigma}$ ) in exactly the 1300 same way as outlined above for  $\Delta\delta$ .

1301 For m, a parameter which is used to create  $\sigma_{
m H}$  by multiplicatively modifying  $\sigma_{
m L}$  (see Equation A4), we specified a normal group-level distribution and estimated both the mean ( $\mu_{m}$ ) and 1302 1303 the standard deviation ( $\sigma_{\rm m}$ ). Because of the exponentiation in Equation A4, values of m below zero 1304 lead to  $\sigma_{\rm H} < \sigma_{\rm L}$ , and vice versa for values above 0. Hence, we placed a normal hyperprior on  $\mu_{\rm m}$  with 1305 a mean of zero. We sought to prevent the fit from favouring extreme differences in slope on the two 1306 sides of the function, as this is against the spirit of the model, which posits a substantial source of 1307 shared noise affecting both sides. Any difference comes from criterial noise that, if too extreme, 1308 would imply regular illogical positioning ( $C_L > C_H$ ) 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  $\sigma_m$  we used a lognormal hyperprior with  $\mu$  of -0.69 and  $\sigma$  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,  $\beta_1$ , to 50, and estimated only the group's modal guess rate ( $\theta_1$ ) which determined the first betadistribution parameter,  $\alpha_1$ , according to:

1318 (A6) 
$$\alpha_{l} = \frac{2\theta_{l} - \theta_{l}\beta_{l} - 1}{\theta_{l} - 1}$$

1319 We used a beta hyperprior on  $\theta_1$  with  $\alpha$  of 1.49 and  $\beta$  of 50. This equates to strongly 1320 expecting a group modal lapse rate around 1%.

## 1321 Multilevel AT-A-GLANCE, differences across conditions.

1322 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 1323 1324 condition. We include accompanying code for this model so readers can see the additions required 1325 to go from 1) single-level, to 2) single-condition multilevel, to 3) multi-condition multilevel model, 1326 which is our final destination. To get to this final model, we still need to specify additional 1327 parameters describing how one or more of our participant-level parameters can vary across 1328 conditions of the experiment. We also need to update our model predictions to incorporate the 1329 effects of these parameters. As noted in the main text methods, this last set of parameters are 1330 conceptually akin to regression coefficients, affecting the model prediction contingent on the value 1331 of the conservative and rebound dummy codes. Dummy codes are 0 or 1 values denoting 1332 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 aboutparticipant, condition and so forth.

1335 The AT-A-GLANCE model envisages participants utilising two criteria to interpret a subjective 1336 difference in arrival times as simultaneous or not. Hence, instructions to be more conservative can 1337 be dealt with by allowing these two criteria to move. However, as previously described, we 1338 reparametrized the criteria as  $\tau$ , their midpoint, and  $\Delta\delta$ , their difference, so it is these parameters 1339 that should be allowed to vary. Each participant therefore required a set of coefficients,  $\beta_{TC}$ ,  $\beta_{TR}$ , 1340  $\beta_{\delta C}$ , and  $\beta_{\delta R}$ , to represent change (compared to baseline). The first subscript represents the 1341 parameter being adjusted and the second represents the <u>Conservative and Rebound conditions</u>. 1342 However, we were mindful that while  $\tau$  is unbounded,  $\Delta\delta$  has a zero lower bound. Hence we allowed 1343 straightforward additive changes to  $\tau$ , but only positive multiplicative ones to  $\Delta\delta$ , with the latter 1344 implemented by exponentiating the relevant coefficient such that positive/negative values translate 1345 to multiplication by greater than or less than 1 respectively. This yields the heterogenous model of 1346 Eqns. 1 and 2 (see main text).

1347 All that now remains to be done for this model is to describe the estimation of group-level 1348 distributions for the experimental effects (the four  $\beta$  coefficients), along with the associated 1349 hyperpriors. For each of these coefficients we specified a normal group-level distribution and 1350 estimated both mean ( $\mu_{\rm m}$ ) and standard deviation ( $\sigma_{\rm m}$ ) parameters (implying eight further group-1351 level parameters). The parameters of these group-level distributions essentially mirror the terms 1352 commonly described as "fixed" and "random" effects (respectively) within a frequentist 1353 general(ised) linear multilevel model framework: The former describe the group-mean effects, the 1354 latter the variation in these effects across participants. We constrain their values with normal 1355 hyperpriors (which due to zero bounding are effectively half-normal for  $\sigma_{\mu}$  parameters) with  $\mu$ s of 0 1356 and  $\sigma$ s of either 80 (for  $\mu_{\tau...}$  and  $\sigma_{\tau...}$  hyperpriors) or 1 (for  $\mu_{\delta...}$  and  $\sigma_{\delta...}$  hyperpriors). To summarise – 1357 we expected zero-size mean effects with zero SD across the group, but modest and even fairly large

1358 effects (and associated variation in effects) would not be unexpected. The final model has 188 1359 parameters (five core plus four  $\beta$  parameters for each of 19 participants, plus nine parameters 1360 describing group variation in core parameters and eight parameters describing group variation in  $\beta$ 1361 parameters). These were estimated based on 513 data points (19 participants x 3 conditions x 9 1362 SOAs).

#### 1363 Multilevel AT-A-GLANCE's alternative account for conservative behaviour

1364 The model described so far can fit simultaneity-judgment data from multiple participants at 1365 once and capture changes across conditions in terms of an adjustment of parameters quantifying 1366 decision criteria. This model essentially represents the hypothesis that simultaneity judgments are 1367 subject to strategic alteration based on these decision criteria. We also created an alternative hard-1368 threshold model variant, in which participants are assumed to maintain their threshold from the pretest but, in the "be conservative" condition, respond "synchronous" on a random subset of trials in 1369 1370 which they actually perceive synchrony. This model essentially represents the hypothesis that what 1371 participants initially report in a simultaneity-judgment task is determined by a structural 1372 thresholding mechanism that does not yield to their subsequent strategic imperatives. This might be 1373 the same gating mechanism underlying multisensory binding/integration if that type of process is 1374 also viewed as all-or-none from a temporal perspective.

1375 The hard-threshold multilevel AT-A-GLANCE model we applied is identical to the multilevel 1376 AT-A-GLANCE model described so far, except in relation to the set of  $\beta$  coefficients used to permit 1377 changes across conditions. Instead of allowing changes to two criteria (in each of two conditions, 1378 relative to the baseline), we now utilise just one change per condition – a proportional reduction in 1379 the number of trials judged synchronous. This can be represented by a pair of coefficients,  $\beta_{\rm C}$  and 1380  $\beta_{\rm R}$ , and yields a heterogenous foundation with a binomially distributed number of "simultaneous" 1381 responses:

1382 (A7)  $S_{X\Delta t} \sim B(30, \beta_X [l + p_{B\Delta t} - lp_{B\Delta t}]),$ 

| 1383 | where $X \in \{B, C, R\}$ and $p_{B\Delta t}$ is defined in Equation 2 (main text, Results).   |  |  |  |  |  |
|------|--|--|--|--|--|--|
| 1384 | For the $eta_{ m C}$ and $eta_{ m R}$ parameters, we modelled variation at the group level as a beta   |  |  |  |  |  |
| 1385 | distribution, but parameterised in terms of a mean parameter:  |  |  |  |  |  |
| 1386 | (A8) $\varphi_{\dots} = \frac{\alpha}{\alpha + \beta}$   |  |  |  |  |  |
| 1387 | And a total count parameter:   |  |  |  |  |  |
| 1388 | (A9) $\lambda_{\dots} = \alpha + \beta$  |  |  |  |  |  |
| 1389 | We followed recommendations in the Stan documentation by specifying hyperpriors that   |  |  |  |  |  |
| 1390 | were beta ( $\alpha$ = 1, $\beta$ = 1, implying uniform) and pareto ( $y_{min}$ = 0.1, $\alpha$ = 0.5) for $\varphi_{}$ and $\lambda_{}$ respectively. |  |  |  |  |  |
| 1391 |  |  |  |  |  |  |
| 1392 | ELA model implementation   |  |  |  |  |  |
| 1393 | Single-level ELA   |  |  |  |  |  |
| 1394 | Our second class of multilevel model built on the four-parameter ELA model (García-Pérez &   |  |  |  |  |  |
| 1395 | Alcalá-Quintana, 2012) which predicts reports of simultaneity as the integral (between two decisior  |  |  |  |  |  |
| 1396 | boundaries) of a difference of exponential distributions. This prediction is described by:   |  |  |  |  |  |
| 1397 | (A10) $P(S \Delta t) = F(\Delta \delta/2; \Delta t) - F(-\Delta \delta/2; \Delta t),$  |  |  |  |  |  |
| 1398 | where function $F$ is given by:  |  |  |  |  |  |
|      |  |  |  |  |  |  |

1399 (A11) 
$$F(d;\Delta t) = \begin{cases} \frac{\lambda_a}{\lambda_a + \lambda_v} \exp[-\lambda_v(\Delta t - \tau + d)] & \text{if } d \leq \Delta t - \tau \\ 1 - \frac{\lambda_v}{\lambda_a + \lambda_v} \exp[\lambda_a(\Delta t - \tau + d)] & \text{if } d > \Delta t - \tau \end{cases}$$

1400 Under this model,  $\lambda_a$  and  $\lambda_v$  are the rate parameters of (shifted) exponential distributions of 1401 arrival times (at the decision centre) for the auditory and visual signals respectively. We have reversed 1402 the sign of García-Pérez and Alcalá-Quintana's  $\tau$  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.

1405 For even further ease of comparison with AT-A-GLANCE, we consider the inverse of the  $\lambda_{\mathrm{a}}$ 1406 parameter  $(\lambda_a^{-1})$ , whose values have a scale and meaning similar to those of AT-A-GLANCE's two noise 1407 parameters. Hence higher values equate to a higher level of internal noise and reduced precision. 1408 Furthermore, in place of  $\lambda_v$  we sampled for m, the natural log of a quantity applied as a multiplier to  $\lambda_a^{-1}$  in order to determine the inverse of  $\lambda_v$ , in a manner analogous to that described in Equation A4 1409 1410 above for AT-A-GLANCE's second noise parameter. Finally, we also included the same lapse parameter 1411 used in our implementation of AT-A-GLANCE, l, such that participants were assumed to make an 1412 (unbiased) guess on a proportion of trials equalling 2*l*. This leads to the following prediction:

1413 (A12) 
$$P(S|\Delta t) = l + (1 - l)[F(\Delta \delta/2; \Delta t) - F(-\Delta \delta/2; \Delta t)],$$

1414 where function *F* is given by:

1415 (A13) 
$$F(d; \Delta t) = \begin{cases} \frac{\exp[m - (\Delta t - \tau - d)/(e^m \lambda_a^{-1})]}{\exp(m) + 1} & \text{if } d \leq \Delta t - \tau \\ 1 - \frac{\exp[m + (\Delta t - \tau - d)/\lambda_a^{-1}]}{\exp(2m) + \exp(m)} & \text{if } d > \Delta t - \tau \end{cases}$$

#### 1416 Multilevel ELA

1417 With both AT-A-GLANCE and ELA utilising a set of largely analogous single-level parameters, 1418 we were able to develop multilevel models of ELA in a very similar way to that outlined above for 1419 AT-A-GLANCE. Hence, we added  $\mu_{ au}$  and  $\sigma_{ au}$  parameters to describe the normal group-level 1420 distribution of  $\tau$ , with Cauchy and lognormal hyperpriors respectively. Similarly, for the lognormal 1421 group-level distribution of  $\Delta\delta$ , we introduced  $\mu_{\delta}$  and  $\sigma_{\delta}$ , with normal and lognormal hyperpriors 1422 respectively, as per the same parameter's treatment in AT-A-GLANCE. The (lognormal) group-level  $\lambda_a^{-1}$  in ELA was dealt with just like the group-level  $\sigma_L$  from AT-A-GLANCE, by including  $\mu_{\lambda a}$  and  $\sigma_{\lambda a}$ 1423 1424 parameters with normal and lognormal hyperpriors respectively. Similarly, we included  $\mu_{\rm m}$  and  $\sigma_{\rm m}$ 1425 to describe the normal group-level distribution of m, with normal and lognormal hyperpriors

1426 respectively, while for l, we added  $\theta_1$  to define the mode of a beta group-level distribution (with a 1427 beta hyperprior). Finally, we added eight parameters to model the means and SDs of the normal 1428 group-level distributions for the four  $\beta$  coefficients which describe changes to  $\tau$  and  $\delta$  across 1429 experimental conditions (for example  $\mu_{\tau C}$  and  $\sigma_{\tau C}$  for the participant-level parameter  $\beta_{\tau C}$  adjusting  $\tau$ 1430 in the conservative condition). These were specified with normal hyperpriors. We also constructed 1431 an alternative model describing the hard-threshold account, with group-level beta distributions of the function multiplier coefficients  $\beta_{\rm C}$  and  $\beta_{\rm R}$ , each described by mean and total count parameters 1432 1433 with beta and pareto hyperpriors respectively, in place of changes to  $\tau$  and  $\Delta\delta$ . In all but a handful of 1434 cases, hyperpriors had parameters exactly as specified for the analogous case in AT-A-GLANCE. The 1435 key exceptions were  $\mu_{\delta}$  and  $\mu_{\lambda a}$ , relating to the distance between criteria and noise for the auditory 1436 stimulus respectively, for which we specified a slightly lower expectation (via setting  $\mu$  = 5.08, with 1437 this default based on 1/5<sup>th</sup> of the range of SOAs). In the case of  $\mu_{\delta}$ , this followed from a 1438 programming choice – we sampled for values of  $\Delta\delta/2$  rather than  $\Delta\delta$ , and hence  $\mu_{\delta}$  should be 1439 around half as large of the equivalent parameter from AT-A-GLANCE. In the case of  $\mu_{\lambda a}$ , estimates 1440 for this parameter from past research tend to be lower than those obtained for  $\mu_{\sigma}$ . 1441 MCD model implementation

## 1442 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  $(S_v(t), S_a(t))$  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*<sub>Corr</sub>). 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:

1449 (A14)  $f_{mod} = t \exp(-t/\tau_{mod})$ 

1450 Where  $f_{mod}$  is an early modality-dependent filter ( $f_a$  and  $f_v$ ) or a late multisensory filter 1451 ( $f_{av}$ ), and  $\tau_{mod}$  is the corresponding filter time constant. The final synchrony estimate is essentially 1452 the time-averaged output formed by multiplying together signals from two units. Each unit 1453 multiplies a single (early) filtered version of one modality with a double (early+late) filtered version 1454 of the other. The final synchrony estimate is then:

1455 (A15)

1456 
$$MCD_{Corr} = \frac{1}{t_2 - t_1} \int_{t_1}^{t_2} ([S_v(t) * f_v(t)] \cdot \{[S_a(t) * f_a(t)] * f_{av}(t)\})$$

1457 
$$\cdot ([S_a(t) * f_a(t)] \cdot \{[S_v(t) * f_v(t)] * f_{av}(t)\}) dt,$$

1458 where \* denotes convolution and  $\cdot$  denotes pointwise multiplication. Finally,  $MCD_{Corr}$  is 1459 used to form binary judgments about synchrony by setting a criterion, with either  $MCD_{Corr}$  itself or 1460 the position of the criterion assumed to be affected by Gaussian trial-by-trial noise, yielding the 1461 prediction:

1462 (A16) 
$$P(S|\Delta t) = \Phi\left(\frac{MCD_{Corr}-C}{\sigma}\right)$$

1463 For our fits, we fixed  $au_{
m av}$  to 786 ms and  $au_{
m v}$  to 87 ms based on fits to other data sets (Parise & Ernst, 2016) and allowed  $\tau_a$ ,  $\sigma$  and C to vary for each observer. We calculated  $MCD_{Corr}$  across a 14 1464 1465 second window centred on the arrival time of the first stimulus (and set to zero except for 10 ms on-1466 off pulses for each signal). We also normalised it by dividing it by the unnormalised MCD<sub>Corr</sub> for a 1467 synchronous input (MCD<sub>Corrs</sub>). This normalisation meant that C could be expected to lie in a range 1468 bounded by 0 and just over 1, and  $\sigma$  should be interpretable on a similar scale. Because Stan does 1469 not currently offer built-in functions for convolution or fast Fourier transformation, we first 1470 determined  $MCD_{Corr}$  for values of  $\tau_a$  from 1 to 200 ms in R, then passed them to Stan as a lookup 1471 table. Within the Stan code, MCD<sub>Corr</sub> values from this table were made continuous (and hence 1472 differentiable) via quadratic interpolation. We also added a lapse parameter, l, consistent with that 1473 applied in our other models:

1474 (A17) 
$$P(S|\Delta t) = l + (1-l)\Phi\left[\frac{\binom{MCD_{Corr}}{MCD_{CorrS}} - C}{\sigma}\right]$$

#### 1475 Multi-level MCD

1476 To upgrade to a multilevel MCD model we dealt with parameter l as per our previous 1477 models, by adding  $\theta_1$  to define the mode of a group-level beta distribution, and specifying a beta 1478 hyperprior on it. For filter time constant  $\tau_a$ , we specified a lognormal group-level distribution and 1479 estimated its two parameters,  $\mu_{\tau a}$  and  $\sigma_{\tau a}$ . For  $\mu_{\tau a}$  we specified a normal hyperprior, with  $\mu$  = 4.34 1480 and  $\sigma$  = 1.09, while the hyperprior for  $\sigma_{\tau a}$  was lognormal with  $\mu$  = -1.39 and  $\sigma$  = 0.25. Together these 1481 correspond to a modal expectation for  $\tau_a$  of around 73 ms. This is comparable to the value of 68 ms 1482 obtained by Parise and Ernst (2016). For criterion C, we specified a normal group-level distribution 1483 and estimated its two parameters,  $\mu_{\rm C}$  and  $\sigma_{\rm C}$ . We gave  $\mu_{\rm C}$  a normal hyperprior, with  $\mu$  and  $\sigma$  both 1484 set at 0.5, and  $\sigma_{\rm C}$  a lognormal hyperprior with  $\mu$  = 0.41 and  $\sigma$  = 1, together implying that C should 1485 have a group mean around 0.5 and SD around 0.55. We then specified a lognormal group-level 1486 distribution for internal-noise parameter  $\sigma$  and estimated both of this distribution's parameters,  $\mu_{\sigma}$ 1487 and  $\sigma_{\sigma}$ . We gave  $\mu_{\sigma}$  a normal hyperprior, with  $\mu$  and  $\sigma$  set at -0.69 and 1 respectively, and  $\sigma_{\sigma}$  a 1488 lognormal hyperprior with  $\mu$  = 3 and  $\sigma$  = 1, together implying that  $\sigma$  was a priori expected to have a 1489 group mode around 0.18, but with larger values remaining plausible. Finally, to allow behaviour to 1490 change across conditions, we allowed criterion C to vary via the introduction of two participant-level 1491  $\beta$  coefficients,  $\beta_{CC}$  and  $\beta_{CR}$ , with multiplicative adjustments to C determined using their exponents. 1492 Each had an associated normal group-level distribution for which we estimated both mean ( $\mu_{C...}$ ) and 1493 standard deviation ( $\sigma_{C...}$ ) parameters (implying four further group-level parameters). Hyperpriors on 1494 these parameters were normal (effectively half-normal for the zero-bounded  $\sigma_{
m C...}$  parameters) with 1495  $\mu$  = 0 and  $\sigma$  = 1. As for our other models, an alternative hard-threshold account was also tested, in 1496 which  $\beta_{CC}$  and  $\beta_{CR}$  were replaced with the function multiplier coefficients  $\beta_{C}$  and  $\beta_{R}$ , each described 1497 by mean and total count parameters with beta and pareto hyperpriors respectively.

1499

## Appendix B: Adequacy of likelihood surface recovery

| 1500 | Before we can consider whether a model is a good description of reality, we need to                   |
|------|---|
| 1501 | determine whether we have successfully explored/characterised the posterior likelihood of the         |
| 1502 | model and its parameters given the data. A model may in principle be perfectly correct, but in        |
| 1503 | practice be impossible to evaluate because of issues such as degeneracy, where parameters trade       |
| 1504 | off so that several different combinations can provide a similarly good fit. Table B1 summarises, for |
| 1505 | the two variants of each of our three models, a set of posterior exploration diagnostics showing how  |
| 1506 | successfully the HMC NUTS algorithm was able to characterise the posterior in each case.              |

1507

| Model                         | < Posterior exploration diagnostics > |              |              |              |
|-------------------------------|---------------------------------------|--------------|--------------|--------------|
|                               | % Divergent iterations                | <b>Max</b> Â | Minimum bESS | Minimum tESS |
| AT-A-GLANCE<br>criterial      | 0.017                                 | 1.045        | 154          | 425          |
| AT-A-GLANCE<br>hard threshold | 0.013                                 | 1.002        | 2324         | 4322         |
| ELA criterial                 | 0.013                                 | 1.002        | 5161         | 6740         |
| ELA hard<br>threshold         | 0.010                                 | 1.002        | 3883         | 5874         |
| MCD criterial                 | 0                                     | 1.001        | 6794         | 11927        |
| MCD hard<br>threshold         | 0                                     | 1.002        | 5177         | 2098         |

1508 Table B1. Posterior exploration diagnostics.

1509

1510 For the AT-A-GLANCE model variant that allowed changes in criteria across conditions, 1511 diagnostics did not completely meet recommendations (Vehtari et al., 2021) despite a relatively long 1512 fit time (around 24 hours per run). In particular, alongside a very small percentage of divergent 1513 iterations, not all parameters reached the ideal level of mixing between chains ( $\hat{R} < 1.01$ ) or the 1514 suggested bulk and tail effective sample sizes (bESS and tESS >400). However, the vast majority of 1515 parameters did meet these recommendations. Furthermore, for the worst offending parameter ( $\mu_{\tau}$ ), 1516 despite a bESS of only 154 the resulting Monte-Carlo standard error (a measure of the precision of 1517 parameter estimation) was just 0.95 (in the context of a mean value of 32.4 ms). The model 1518 predictions also mapped well onto the data (see main text Results). We therefore believe that from 1519 a practical point of view, this model was characterised adequately to allow us to make sensible 1520 comments regarding how well it described the data compared to other models explored here. 1521 For the second, hard-threshold, variant of the AT-A-GLANCE model, exploration diagnostics 1522 met all recommendations with the exception of a very small percentage of divergent iterations. The 1523 posterior exploration diagnostics from Table 1 also indicate that both variants of both ELA and MCD 1524 models met recommendations in terms of chain mixing and bulk and tail effective sample sizes, with 1525 only a very small percentage of divergent iterations (<=0.013%). This suggests that the HMC NUTS 1526 sampling algorithm was able to properly characterise the posterior in each case. The ELA model's

1527 posterior proved particularly easy to characterise, with fits returning in under 30 minutes for these

1528 data. Parameter recovery was assessed separately (via simulation) – see Appendix D.
## Appendix C: Assessment of Bayesian design choices

1531 In implementing Bayesian multilevel models, we had to make various decisions, including 1532 specifying the distributions with which we expected participant-level parameters to vary across the 1533 group. We also had to set prior expectations for the parameters of these group-level distributions, 1534 known as hyperpriors. In Figure C1, we consider these choices for the AT-A-GLANCE model variant 1535 permitting changes in decision criteria across conditions. The smaller graphs within Figure C1 1536 illustrate the posterior distributions obtained. They focus on the subset of group-level parameters 1537 relating to behaviour in the baseline condition (parts a-e). We also illustrate two of the remaining 1538 eight group-level parameters that relate to changes in behaviour in the conservative condition 1539 (specifically changes in the width of the simultaneity function; part f). Hyperpriors (plotted only 1540 across the limited range required to characterise the posteriors) are shown for comparison (dashed 1541 lines). Posteriors are markedly less diffuse than hyperpriors, rarely coincide with their modes, and 1542 don't appear to have been constrained by their edges. It is thus clear that posteriors were not 1543 unduly influenced by our choices regarding hyperpriors, and must have been very largely 1544 determined by the data. The figure also illustrates how these group-level estimates in turn 1545 parameterise group-distributions (shown as black against red lines in larger graphs). These describe 1546 how participant-level parameters vary across the group. They can be compared with the model's 1547 participant-level estimates for each individual (shown as circles) and a kernel-density plot derived 1548 from them (shown opposite parametric predictions). In general, the choices of distributions seem 1549 reasonable, although the participant-level estimates will have been constrained by these choices 1550 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

1558 participant-level parameters are shown as black circles, their mean as a solid horizontal line, and a 1559 kernel-density estimate based on these estimates completes the hourglass plot as the right-hand 1560 lobe. (a) Group-level parameters  $\mu_{\tau}$  and  $\sigma_{\tau}$  which describe the (normal) distribution of the 1561 participant-level parameter  $\tau$ . This in turn describes the central tendency of the simultaneity function. 1562 (b) Group-level parameters  $\mu_{\delta}$  and  $\sigma_{\delta}$  which describe the (lognormal) distribution of the participantlevel parameter  $\Delta\delta$ . This in turn describes the width of the simultaneity function. **(c)** Group-level 1563 1564 parameters  $\mu_{\sigma}$  and  $\sigma_{\sigma}$  which describe the (lognormal) distribution of the participant-level parameter 1565  $\sigma_{\rm L}$ . This in turn describes the (inverse) slope of the simultaneity function's left flank. (d) Group-level 1566 parameters  $\mu_m$  and  $\sigma_m$  which describe the (normal) distribution of the participant-level parameter 1567 m. The exponent of m is multiplied by  $\sigma_{\rm L}$  in order to yield the (inverse) slope of the simultaneity 1568 function's right flank. Hence the group-mean value illustrated here implies  $\sigma_{\rm H}$  was typically around 1569 1.3 times as large as  $\sigma_L$ . (e) Group-level parameter  $\vartheta_l$  which fixes the mode of the (beta) distribution 1570 of the participant-level parameter *l*. This in turn describes the (half) lapse rate defining lower/upper 1571 bounds on the simultaneity function. (f) Group-level parameters  $\mu_{\delta C}$  and  $\sigma_{\delta C}$  which describe the 1572 (normal) distribution of the participant-level coefficient  $\beta_{\delta C}$ . This is in turn exponentiated to form a 1573 multiplier quantifying how  $\Delta\delta$  changes in the "be conservative" condition of the experiment (see 1574 main text Results section for further details relating to interpreting this coefficient).

## **Appendix D: Parameter recovery simulations**

1577 To check that our methods were capable of adequately recovering model parameters, we 1578 simulated our experiment. We drew binomial-distributed random responses based on model-1579 predicted "proportion judged synchronous" values for 19 participants in three conditions each with 1580 9 SOAs and 30 trials per SOA. This was done based on known parameter values for each of our three 1581 (criterial-variant) models. We drew these parameter values at random from distributions that 1582 approximated those we had estimated for the population when fitting the models to our actual data. 1583 For example, when assessing parameter recovery for the AT-A-GLANCE model, the  $\tau$  parameter for 1584 each simulated participant was drawn from a distribution similar to that shown as black against a red 1585 background in the hourglass plot of Appendix C Figure C1a, and so on for other parameters. 1586 Simulated data were then fit using the model that had generated them via a slightly truncated 1587 version of the same procedure that we applied to real data for our main analyses (with 5000 rather 1588 than 10000 post warmup samples per chain, to reduce computation time). 1589 Figures D1 to D3 show actual vs. recovered parameter values (alongside the ideal lines of 1590 equality) for the criterial AT-A-GLANCE, ELA, and MCD models respectively. Parameters are in 1591 general recovered fairly well based on the numbers of trials and fitting procedures used in our 1592 experiment.



1593

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)**  $\beta$  Model parameters describing changes in position ( $\tau$ ) and width ( $\Delta\delta$ ) of the psychometric function in the <u>C</u>onservative and <u>R</u>ebound conditions.



1601

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 ( $\tau$ ), width ( $\Delta\delta$ ), shape ( $\lambda_a^{-1}$  and m), and (half) lapse rate (l). (*f-i*)  $\beta$  Model parameters describing changes in position ( $\tau$ ) and width ( $\Delta\delta$ ) of the psychometric function in the <u>C</u>onservative and <u>R</u>ebound conditions.



1610Figure D3. Parameter recovery simulation with criterial MCD as the generative model. Dashed black1611line indicates equality for generative and recovered parameters; r = Pearson correlation coefficient.1612(a-d) Model parameters describing baseline performance. These affect the psychometric function via1613the model's auditory-filter time constant ( $\tau_a$ ), decision criterion (C), noise ( $\sigma$ ), and (half) lapse rate1614(l). (e-f)  $\beta$  Model parameters describing changes in the decision criterion (C) affecting the1615psychometric function in the <u>C</u>onservative and <u>R</u>ebound conditions respectively.

## 1617 Appendix E: Ability of PSIS-LOO metric to compensate model complexity and discriminate true

## 1618 from false models

1619 Appendix D, above, describes how we used each of our three (criterial-variant) models to 1620 create a simulated data set and fit that data set with the generative (i.e. true) model in order to 1621 assess parameter recovery. Further to this, we additionally recorded PSIS-LOO as a measure of 1622 goodness of fit (as per our main data analysis, but without additional leave-one-out substitution for 1623 Pareto ks > 1.0 to reduce computation time; hence a somewhat noisier approach to goodness-of-fit 1624 estimation). We then fit both of the alternative (i.e. false) models to that same simulated data and 1625 recorded PSIS-LOO for them in the same way. Finally, we repeated the whole procedure for a second 1626 run.

1627 The resulting PSIS-LOO values are shown in Table E1. AT-A-GLANCE and ELA have identical 1628 numbers of free parameters. AT-A-GLANCE yields higher values of PSIS-LOO compared to ELA when 1629 it is the generative model (as expected). PSIS-LOO is more similar between these models when ELA is 1630 generative, although ELA wins (significantly) on one of the two runs. These results suggest that AT-A-1631 GLANCE may be better able to mimic ELA than vice versa, at least with our procedures. The MCD 1632 model has less free parameters than both AT-A-GLANCE and ELA. As PSIS-LOO is intended to 1633 estimate goodness of fit while taking appropriate account of model complexity, MCD should 1634 nonetheless outperform the other two models when it is generative. It indeed scores significantly 1635 better, suggesting that the PSIS-LOO metric is working as intended in the current context and 1636 favouring a parametrically simpler generative model over more complex (but false) alternatives.

1637

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</li>
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)   |