Auditory and visual durations load a unitary working-memory resource

Kielan Yarrow[1*], Carine Samba[1], Carmen Kohl[2] & Derek H. Arnold[3]

[1] *Department of Psychology, City, University of London*

[2] *Department of Neuroscience, Brown University*

[3] *School of Psychology, The University of Queensland*

Short title: A unitary working-memory resource for duration

* Author for correspondence:

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

Tel: +44 (0)20 7040 8530
Fax: +44 (0)20 7040 8580
Email: kielan.yarrow.1@city.ac.uk

Abstract

Items in working memory are typically defined by various attributes, such as colour (for visual objects) and pitch (for auditory objects). The attribute of duration can be signalled by multiple modalities, but has received relatively little attention from a working-memory perspective. While the existence of specialist stores (e.g. the phonological loop and visuospatial sketchpad) is often asserted in the wider working-memory literature, the interval-timing literature has more often implied a unitary (amodal) store. Here we combine two modelling frameworks to probe the basis of working memory for duration; a Bayesian-observer framework, previously used to explain behaviour in duration-reproduction tasks, and mixture models, describing distributions of continuous reports about items in working memory. We modelled different storage mechanisms, such as a limited number of fixed-resolution slots or a resource spread between items at a cost to resolution, in order to ask whether items from different sensory modalities are maintained in separate, independent stores. We initially analysed data from 32 participants, who memorised between one and eight items before reproducing the duration of a randomly selected target. In separate blocks, items could be all visual, all auditory, or an alternating mixture of both. A small control experiment included a further condition with pre-cuing of target modality. Certain kinds of slot models, resource models, and combination models incorporating both mechanisms could account for the data. However, looking across all plausible models, the decline in performance with increasing memory load was most consistent with a single store for event durations, regardless of stimulus modality.

Keywords: working memory, duration, interval timing, unimodal, multimodal

1. Introduction

The term working memory describes a limited-capacity process (or set of processes) underlying the retention and manipulation of information over a short timescale. Working memory likely supports myriad psychological functions, and may even form the stuff of consciousness (Baars, 2005). Classic models of working memory posit several functionally distinct memory stores, each with strong links to a particular sensory modality (Baddeley, 2012).

Traditional measures of working memory (e.g. digit span) assess knowledge about categorically distinct stimuli, but the inputs we encounter daily can differ on a variety of attributes (e.g. colour, orientation, position, form) which may or may not be easily categorised, and which may or may not be precisely represented in working memory. One attribute that has received relatively little attention from a working-memory perspective is the duration of a stimulus. Interestingly, an amodal working memory store, functionally insulated from the wider working-memory network, is often invoked in specialist models of duration perception (e.g. in scalar expectancy theory; Gibbon, Church, & Meck, 1984). Yet the amodal nature of this store remains in question.

Here, we examine how stimulus durations are represented in working memory. We ask whether or not visual and auditory durations are maintained in separate stores that tap independent resources (Bratzke, Quinn, Ulrich, & Bausenhart, 2016; Bratzke & Ulrich, 2020; Rattat & Picard, 2012). Because we address this question via a model-based analysis, en route to promoting an answer we also consider whether the storage mechanism for perceived durations can plausibly be described as a limited set of high-fidelity memory slots, and/or as a continuous resource that can be divided between stored items (Ma, Husain, & Bays, 2014). These supplemental questions are interesting in their own right, but here they primarily serve to demonstrate that our conclusions, regarding the

existence of a unitary memory store for durations, can generalise beyond a single (assumed) storage format. We expand on both our core and supplemental questions below.

1.1 Separate memory stores for the duration of auditory and visual stimuli?

There is some debate regarding whether the functional architecture supporting human perception of duration should be considered centralised (and therefore amodal) or distributed (Ivry & Schlerf, 2008; Paton & Buonomano, 2018). For example, focussing on behavioural evidence, interval timing has broadly similar psychophysical properties in different sensory modalities, and training in one modality, which often leads to improvements in the precision of timing, can transfer to a second modality (Bartolo & Merchant, 2009; Bratzke, Seifried, & Ulrich, 2012; Nagarajan, Blake, Wright, Byl, & Merzenich, 1998, but see Lapid, Ulrich, & Rammsayer, 2009). These sorts of findings have encouraged researchers to conclude in favour of a centralised clock.

However, other behavioural evidence suggests that stimulus durations might be processed by modality-specific mechanisms (e.g. Ball, Arnold, & Yarrow, 2017; Hartcher-O'Brien, Di Luca, & Ernst, 2014; Heron et al., 2012). For example, Ball et al. (2017) found an improvement in precision for visuotactile bimodal stimuli, compared to unimodal stimuli from either modality. This finding implies that independent sources of noise might limit the precision of tactile and visual duration judgements, such that averaging can improve performance in the bimodal case. The limiting noise was shown to relate to the processing of duration, rather than to the mere registration of events that mark the onset and offset of an interval, but sadly this study could not further isolate the locus of the noise. Hence, the independent modality-specific sources of noise that dominated decisions could have arisen at either a sensory level (i.e. during temporal accumulation) or later, within a multi-faceted working memory system (i.e. within separate modality-specific stores).

Experimentalists have mounted various investigations regarding how *long-term* memory representations of duration emerge when intervals are presented in more than one modality (e.g. Filippopoulos, Hallworth, Lee, & Wearden, 2013; Ogden, Wearden, & Jones, 2010; Roach, McGraw, Whitaker, & Heron, 2017). However, to our knowledge the idea of separate modality-specific *working-memory* stores for duration has been probed less. It was addressed in the most specific manner so far by Rattat and Picard (2012), who sought and obtained a double dissociation of modality-selective interference effects during retention of an interval's duration. Participants held an interval in mind for eight seconds while performing a secondary task, then compared it against a target using a same/different judgment. Visuospatial tracking was found to disrupt only visual duration retention, whereas articulatory suppression disrupted only auditory duration retention. This finding is highly suggestive of the existence of separate, auditory and visual memory stores for duration. However, Bratzke et al. (2016) have replicated this study with very minor methodological changes, and obtained a contrasting result, finding evidence for only auditory task interference, which affected interval timing for both modalities. Furthermore, in a subsequent study with different secondary tasks (tone and colour discrimination) and a bidirectional analysis of interference effects, the same group found some evidence of selective auditory interference, but no visual interference (Bratzke & Ulrich, 2020). Hence dual-task research on this topic appears inconclusive at the present time.

There is, however, an alternative to asking if standard visual and auditory interval timing tasks are negatively affected by a very different secondary task (such as articulatory suppression). This alternative becomes available if memory is loaded with several items. In this situation, one can ask whether presenting a mixture of stimuli (which would permit use of multiple stores, if they are available) can enhance performance relative to maintaining the same overall number of stimuli from a single modality. For example, asking people to remember the durations of two visual and two

auditory inputs should be easier, compared to a set of four visual inputs, if they are maintained in at least partially independent single-modality stores. This logic is sometimes reframed, to predict impaired (rather than enhanced) performance when items are added from a second category (which would increase memory load, but only if these items tap the same unitary store).

This general approach has previously been applied with, for example, different categories of visual object (e.g. faces, body postures, cars etc.; Endress, Korjoukov, & Bonatti, 2017; Wong, Peterson, & Thompson, 2008; Wood, 2007). However, of particular relevance here are reports investigating working memory for both visual and auditory objects (e.g. coloured squares and spoken numbers, or white dots and bird calls; Cowan, Saults, & Blume, 2014; Fougnie, Zughni, Godwin, & Marois, 2015; Uittenhove, Chaabi, Camos, & Barrouillet, 2019). The work of Cowan et al. (2014) is illustrative of this approach. In a series of experiments, they found only limited impairment when participants had to memorise five visual and five auditory items, compared to focussing on just one or the other set. The authors used a slot model (see below) to interpret results, implying a weak central/overlapping resource supplemented by two larger-capacity independent unimodal stores.

We are not, however, aware of any equivalent work specifically assessing memory for the duration of stimuli. The closest work we have found is an experiment by Gamache and Grondin (2010), who had participants remember either one or two intervals that could be both visual, both auditory, or one of each. Unexpectedly, they found no decrement with increasing load from one to two items *except* in the AV case – a result which does not sit easily with either single or dual-store accounts. However, they presented stimuli in a temporally overlapping format, which complicates the task and its interpretation considerably (see e.g. Bryce, Seifried-Dübon, & Bratzke, 2015; van Rijn & Taatgen, 2008, for related work). In order to utilise a load-based approach (like Cowan et al., 2014) to address the question of multiple working memory stores for duration, we must simplify this aspect of the procedure, so that each item can in principle be unambiguously encoded into memory. We must

also consider models of working memory under load, and how they should be adjusted to deal with duration as a stimulus attribute, an issue to which we turn next.

1.2 Slot vs resource models of working memory

One common approach in the working-memory literature is to use all-or-none behavioural assessments (e.g. digit recall or change detection) in conjunction with easily categorised stimuli (e.g. digits, nameable colours). This tradition has provided an important source of support for the view that working memory contains a limited number of *slots* into which items can be allocated, to be retained for short periods with near-perfect resolution. When all slots are filled, additional items cannot be stored. The idea of a limited item capacity for working memory is an old one (Miller, 1956), that has often been used to characterise behaviour with a single number (representing memory capacity) that can be compared across experimental tasks (e.g. Cowan, 2001; Luck & Vogel, 1997).

Slot models can be contrasted with the idea of a more continuous resource, that may be divided between any number of items, at a cost to resolution (Ma et al., 2014; Palmer, 1990). This idea has become more widely appreciated since the development of paradigms that use continuous measures, such as pointing to a colour on a colour wheel, to index the contents of memory. This approach facilitates an analysis comparing the full distribution of responses that differ along some circular attribute (e.g. colour, orientation, motion direction) to the distribution predicted by a model (Bays & Husain, 2008; Wilken & Ma, 2004; Zhang & Luck, 2008). For example, a simple slot model predicts that responses should be a mixture of a uniform distribution (composed of guesses, when memory capacity has been exceeded) and a Gaussian distribution (reflecting sensory/memory noise when items are correctly recalled). By contrast, a simple resource model (e.g. one without binding

errors, or favoured item positions) predicts only a Gaussian distribution (as all items are present in memory) but with an increasing width, reflecting increased representational noise as memory load increases. Of course, slot-based and resolution-based capacity limits might co-exist, and perhaps relate differently to wider cognitive functioning (Fukuda, Vogel, Mayr, & Awh, 2010).

We know of only a handful of papers which have substantially varied item load / set size, while assessing memory for duration (Fan & Yotsumoto, 2018; Manohar & Husain, 2016; Teki & Griffiths, 2014; Teki & Griffiths, 2016). Of these, most have used a method that is, in principle, amenable to an analysis of continuous response distributions: Interval reproduction. Teki and Griffiths (2014; see also Teki & Griffiths, 2016) presented a series of clicks that demarcated one to four empty intervals of varying duration. Similarly, Manohar and Husain (2016) presented one to five tones varying from 200 to 2000 ms (separated by 500 ms silent periods). In both studies participants attempted to reproduce a single target interval, and precision dropped with increasing load, suggesting a memory limit of some kind. However, no formal modelling of the distribution of responses was applied in either study.

A possible reason for this lack of formal mixture modelling becomes apparent when one considers the nature of duration as an attribute. Unlike attributes such as colour and orientation, duration cannot easily be expressed in a circular space (i.e. one that wraps around, as when orientations of 360 degrees wrap to zero degrees). Such spaces simplify the predicted mixture distributions from slot models, because so long as no part of the stimulus space is favoured during stimulus selection, a correctly retained item will be reproduced without bias, while guessing should be at random (relative to the target stimulus) and thus uniform. By contrast, for a non-circular stimulus space, such as duration, it will always be possible to determine a central tendency in the stimulus set. This means that mixture models must be adapted to apply to this kind of situation.

Fortunately, modelling work addressing performance in interval reproduction tasks offers the requisite framework for combination with slot and resource models of working memory. Such combined modelling offers a principled means of assessing whether working memory resources for duration overlap between modalities. An evolutionarily optimised observer, for whom the impression of a stimulus is imperfect as a result of neural noise, could use this knowledge via a process of Bayesian inference (essentially averaging what is sensed with what is expected, with weightings that reflect the reliability of each source of information). This would improve performance when performing a duration recollection task according to widely accepted criteria, such as minimising the squared error of responses. In the case of duration reproduction, such a Bayesian-observer model has been shown to be plausible, providing, for example, a good account of Vierordt's law – the classically observed overproduction of short intervals and underproduction of longer intervals (Acerbi, Wolpert, & Vijayakumar, 2012; Cicchini, Arrighi, Cecchetti, Giusti, & Burr, 2012; Jazayeri & Shadlen, 2010). These authors have illustrated and explained both this central tendency in duration reproductions, and a tendency to show greater central bias for long relative to short duration stimuli, which is in line with the greater scalar noise present in the stimulus (alongside an identical prior expectation). However, models have not yet been extended to deal with situations where more than a single stimulus is encoded for subsequent reproduction. A Bayesian-observer model could therefore be integrated into models of working memory, to provide quantitative measures that describe memory resources and their degree of overlap between modalities.

1.3 The current study

Our brief review of relevant literature highlights two gaps in knowledge. First, and of primary interest, it is unclear if separate working memory stores maintain duration estimates for events from different sensory modalities (e.g. vision and audition), as no studies have addressed this specific

question by loading memory with more than two items. Second, it is unclear whether working memory for duration is best described by a slot model, a resource model, or by some combination of both. Here we use a well-validated Bayesian-observer model, to extend mixture modelling approaches to a non-circular stimulus attribute (duration). This allows us to compare several putative models, and, from the most plausible models, assess whether independent memory stores might exist.

To provide appropriate data, we had participants reproduce one from a sequence of up to eight visual or auditory durations, with sequences being either entirely unimodal, or comprising memorised items that alternated between modalities. In addition to validating established results (for example the central tendency bias in interval reproductions, and the greater precision of auditory vs. visual timing) we tested which of our two hypotheses regarding working memory for duration was best supported. If independent working-memory stores exist, we expected performance in a given intermixed condition to be similar to that observed in the corresponding unimodal condition with *half* the memory load (as items can be shared between two stores in the intermixed case). If only a single store is being used, we expected performance in intermixed conditions to be similar to that observed in unimodal conditions with the exact same memory load.

Predictions about how any particular performance metric should vary under our contrasting hypotheses are likely to be model dependent, e.g. to reflect the principles by which working memory for duration operate within a single modality. Hence, rather than relying on our intuitions regarding expected patterns across conditions for a particular arbitrary performance measure, such as error scores or percentage correct, we have applied formal modelling. Specifically, we introduced a theoretically meaningful free parameter, estimated for each of eight putative models. This model parameter acted to divide the memory load experienced in intermixed conditions (e.g. four items) to provide an effective load (e.g. two items, if the parameter took a value of 2). Hence, the value of this

free parameter could be used to assess the degree of divergence of data from each contrasting hypothesis, with a single-store account predicting a parameter value of 1, and an independent-store account predicting a value of 2. Because this process is somewhat involved, we present further details toward the end of the methods section, once a fuller account has been provided regarding model implementation.

2. Material and Methods

2.1 Participants

We recruited 50 participants. Of these, 45 were recruited as the main experimental group, sampled opportunistically via either a first-year undergraduate psychology participant pool (receiving course credit in exchange for participation) or recruited by word of mouth (with some receiving £8 per hour for participating). To verify task engagement, participants were only included in the final analysis if there was a significant (one-tailed $p < 0.05$) correlation between stimulus durations and their reproduced durations on trials with a memory load of one item (see design, below). This criterion, which was established prior to data collection, yielded a final sample of 30 participants[1] (23 female, mean age 22.5, SD 6.7). We tested two further participants, hereafter termed "observers", more extensively (one author, and one observer who was naïve as to the experimental hypotheses, both male, ages 42 and 26 respectively). Their data were used to illustrate model fits (see results, below). Finally, we recruited three further naïve participants, who, along with the aforementioned author (three out of four female, mean age 31.5, SD 9.2) completed a control experiment. The experiments

[1] One participant (with an r of 0.228, against a critical r value of 0.195) was erroneously rejected and replaced during recruitment, but was not added back following discovery of the error, as doing so would have compromised counterbalancing.

were approved following the procedures of the local research ethics committee. All participants provided written informed consent.

2.2 Apparatus and stimuli

A PC running Windows and custom Matlab R2013b software (with the Psychtoolbox extension; Brainard, 1997) was used to control the experiment. Two digital buttons were pinched lightly, one between each forefinger and thumb, and interfaced via a national instrument X-series PCIe-6323 A/D card sampling at 100,000 Hz to enable participants to reproduce stimuli. The duration of each target item was drawn at random from a uniform distribution, ranging from 400 to 1000 ms. On each trial, up to eight visual memory items could be displayed at the centre of a Cambridge Research Systems (CRS) Display ++ M0623 monitor (vertical refresh rate 100 Hz), coloured red, green, blue, yellow, white, turquoise, purple or grey, with colours always appearing in that order. A Gaussian transparency window, with a standard deviation subtending approximately two degrees visual angle, created the appearance of blobs fading into a black background. Other visual stimuli (text containing instructions/feedback, and a fixation point) were presented in white. Auditory memory items (pure tones of 400, 600, 800, 1200, 1600, 2400, 3200 or 4800 Hz, generated at 44.1 kHz) were played through Sennheiser PX360 headphones at a comfortable intensity, always in an ascending order of pitch.

2.3 Design

Two factors, sequence modality (auditory, visual, or intermixed) and memory load (1 vs. 2 vs. 4 vs. 8 items) were factorially combined in a 3x4 repeated-measures design. Modality was blocked, with block order counterbalanced across the 30 participants. The sequence of different loads was randomised within blocks. Each block contained 96 trials (24 repetitions of each load). Participants

completed a short practice block, containing 24 trials with load 1 (from their starting modality condition) followed by several warmup trials from the full procedure (until they had experienced different loads and reported full understanding of the task) followed by one block per condition, all over the course of a single one-hour experimental session. By contrast, the two additional observers completed three blocks per condition, across three one-hour sessions. For them, each session contained one block per condition, with order of conditions varying randomly across sessions (but no systematic counterbalancing).

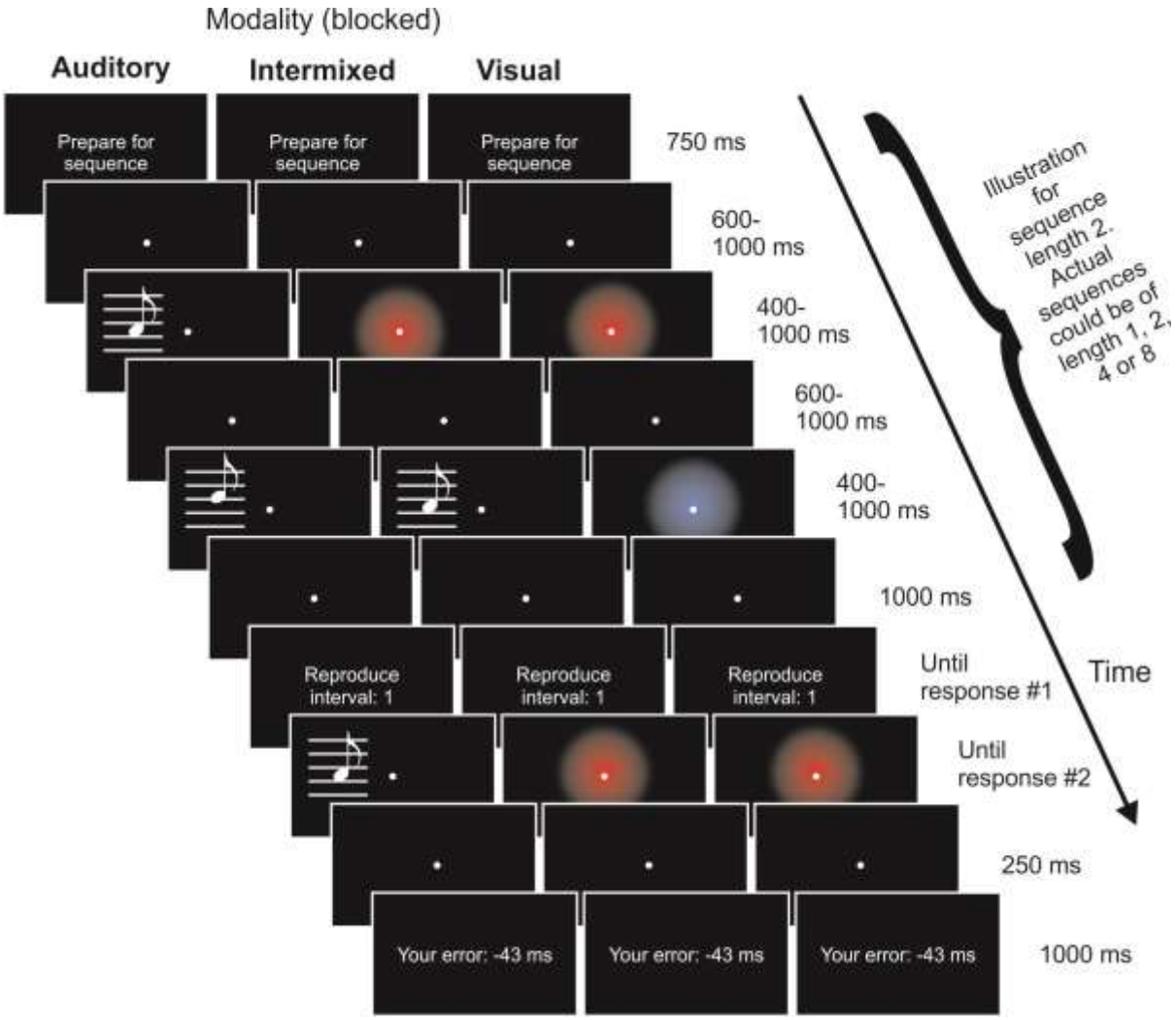

Figure 1. Schematic of the experimental task and design. A trial is shown in which memory is loaded with two items, but load was randomly varied within each block (between 1, 2, 4 or 8 items). The white musical notes represent auditory stimuli and were not shown on screen.

2.4 Procedure

The procedure is schematised in Fig. 1. After a prompt to prepare for the upcoming sequence (750 ms) the screen alternated between a fixation dot (600-1000 ms) and a memory item (400-1000 ms) between one and eight times. In visual conditions, all memory items were visual. In auditory conditions, all memory items were auditory. In intermixed conditions, visual and auditory items alternated, beginning with a visual or auditory stimulus at random. Following the presentation of all memory items, a 1000 ms fixation dot preceded a prompt indicating which item was the target to reproduce (with all possible item positions selected equally often as targets for each load across a complete block of trials). Participants pinched one button to initiate their reproduction, causing the target stimulus (either visual or auditory) to reactivate. Participants then pinched the other button in order to end their reproduction, causing the target stimulus to desist. After 250 ms, a 1000 ms feedback screen informed participants of their reproduction error (in milliseconds) and a new trial began.

2.4.1 Control Experiment with pre-cuing

For the control experiment, a fourth condition was added to the sequence-modality factor, to create a 4x4 repeated-measures design. The new condition was identical to the intermixed condition, except that the prompt (now on screen until a button press rather than for a fixed 750 ms) included a pre-cue informing the participant which modality would contain the target, thus allowing them to focus on just one subset of stimuli. These cued intermixed trials were included as a positive control in order to provide an opportunity for participants to strategically double their capacity (by halving the number of stimuli that had to be attended). On non-cued trials, the prompt informed participants whether the upcoming trial would include all visual, all auditory, or a mixture of stimuli from both modalities. Unlike the main experiment, where sequence modality was blocked, all 4x4 conditions were now randomly interleaved in a block of 256 trials, taking around an hour to complete. Each participant completed three such blocks, typically on separate days.

2.5 Descriptive analysis

For each participant, raw data consisted of target durations and their corresponding reproduced durations, sorted by load, modality of sequence, and modality of target (for intermixed sequences). Data were initially trimmed (separately for each participant) to remove button errors, by fitting each load/sequence-modality condition with a linear regression and iteratively removing outliers (i.e. repeating this fit and trim process, potentially several times) until all studentised residuals fell within +/- 3.09 (a procedure which should exclude only around 0.1% of non-erroneous data in either tail).

To permit an informal preliminary analysis, trimmed durations were converted to errors (i.e. reproduced minus target durations) and squared, before averaging and square rooting to yield root mean squared error (RMSE) values for each condition. This approach is conceptually similar to that employed by Teki and Griffiths (2014), although they used absolute error, and inverted scores to yield a precision-like measure. Although RMSE can be decomposed further, into constant and variable error, doing so under the current design requires recourse to a descriptive model of behaviour, such as linear regression (cf. Manohar & Husain, 2016). As we planned to investigate our data using more psychologically meaningful (and complete) models (outlined below), we do not report such an (intermediate/descriptive) decomposition.

2.6 Modelling

2.6.1 Modelling (1): General

For our formal analyses, several candidate models were fitted to trimmed data using Matlab (The Mathworks, Natick, U.S.A) in order to recover a set of parameters that maximise the likelihood of each model given the data. Parameter values were adjusted using a differential evolution algorithm (Price, Storn, & Lampinen, 2006). The resulting (log) likelihoods were converted to Akaike

Information Criteria and Bayesian Information Criteria (AIC and BIC) via standard formulae. Both AIC and BIC include a punishment term for additional model parameters, because more complex models have a greater capacity to capture the noise that accompanies data, and thus will tend to fit data better in absolute terms. Information criteria offer a relatively simple approach that still permits comparison of non-nested models, like those used here. This model comparison addressed whether the storage mechanism could plausibly be described as a limited set of high-fidelity memory slots, and/or as a continuous resource that can be divided between stored items. As outlined further below, the value of one of the model parameters was used to address our key question: Whether auditory and visual durations were maintained in separate working-memory stores drawing on independent neural resources. Models are summarised in Table 1.

Table 1. Parameter summary for eight models of duration reproduction under memory load. See main text for description of how these parameters come to affect performance across the 3x4 experimental conditions. The first four models incorporate either a slot-limit or a continuous resource leading to variable memory noise, while the final four models cross-combine these features.

Model name	**Number of parameters**	**Parameters handling single-item performance: Weber ratios.**	**Parameters handling load-induced performance decrements: Capacity and/or noise exponents.**	**Parameter handling changes to performance in intermixed conditions: Load denominator**
Slots plus guess from prior	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis}	d
Slots plus guess at mean of prior	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis}	d
Memory noise with late Bayes integration	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	n_{aud} , n_{vis}	d
Memory noise with early Bayes integration	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	n_{aud} , n_{vis}	d
Guess from prior, combined with late Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d
Guess from prior, combined with early Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d
Guess at mean of prior, combined with late Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d
Guess at mean of prior, combined with early Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d

2.6.2 Modelling (2): Interval reproduction without memory load

We first outline the interval reproduction model previously developed by others in order to account for the reproduction of a single item (Jazayeri & Shadlen, 2010). We then progress to describing

additional model parameters, which differentiate this core Bayesian model into a set of candidate models each capable of describing duration reproductions, but assuming distinct modes of working memory storage. Our models were implemented in code which is freely available at DOI 10.25383/city.11842410.

Starting with the single-item case, all models are Bayesian observer models of a form introduced by Jazayeri and Shadlen (2010) to explain the central tendency bias observed in duration reproduction (known as Vierordt's law). We simulated the basic architecture illustrated in Fig. 2.[2] Any given stimulus duration (D_s) is initially combined with a source of stimulus noise, assumed to be a zero-mean Gaussian. The first two model parameters represent auditory and visual sensory Weber ratios (W_s , e.g. W_{sAud}) and could take on values between 0.01 and 1. Because the precision of interval timing is known to decrease with increasing stimulus duration, according to the scalar property (a particular case of Weber's law; Wearden & Lejeune, 2008), the relevant sensory Weber ratio was multiplied with the stimulus duration to determine the standard deviation of Gaussian noise (D_sW_s). Such scalar noise has previously been shown to provide a better account of human behavioural data (compared to constant noise) for the current class of models (Acerbi et al., 2012). Simulation then proceeded through implementation of a Bayesian estimator, which first inferred a likelihood function from noise-corrupted sensory estimates (D_{sn}), based on an accurate knowledge of underlying sensory noise:

$$L(D_s|D_{sn}) = \frac{1}{D_sW_s\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{D_{sn}-D_s}{D_sW_s}\right)^2} \quad (1)$$

[2] To reduce both computation time and simulation noise, in place of random sampling for each source of noise, we divided the probability space from 0.01 to 0.99 in 50 steps of 0.02, and applied an inverse Gaussian function to these values to recover pseudo-simulated noise scores. This process distorts model predictions relative to a true Monte Carlo simulation, but informal explorations suggested this distortion was negligible with parameter values similar to those that we recovered.

This likelihood function was combined with a prior according to Bayes law (i.e. the likelihood was multiplied by the prior, then normalised to yield a valid posterior probability density function). However, the manner in which the brain represents prior distributions remains to be fully elucidated. Jazayeri and Shadlen (2010) assumed a uniform density function, which reflected the experimental stimuli, but subsequent work has indicated that Bayesian models of duration reproduction better match human behaviour when the prior is assumed to be a Normal approximation of the truly uniform distribution of stimuli (Acerbi et al., 2012; Cicchini et al., 2012), here:

$$P(D_s) = \varphi\left((D_s - 0.7)\left(\frac{0.6}{\sqrt{12}}\right)\right) \quad (2)$$

Where φ is the standard normal probability density function. Hence we adopted this approach.

Following the combination of the prior with the sensory likelihood, a perceptual point estimate was determined as the mean of the posterior (i.e. the Bayes least squares model of Jazayeri and Shadlen, 2010; see also Acerbi et al., 2012). This optimises perceptual estimation based on a squared-error cost function. Finally, the reproduction of the perceptual estimate was modelled with an additional source of scalar Gaussian (zero-mean) noise, with a standard deviation obtained by multiplying the perceptual estimate with a modality-appropriate motor Weber ratio (W_m). For this purpose, the third and fourth model parameters represented auditory and visual motor Weber ratios, respectively.

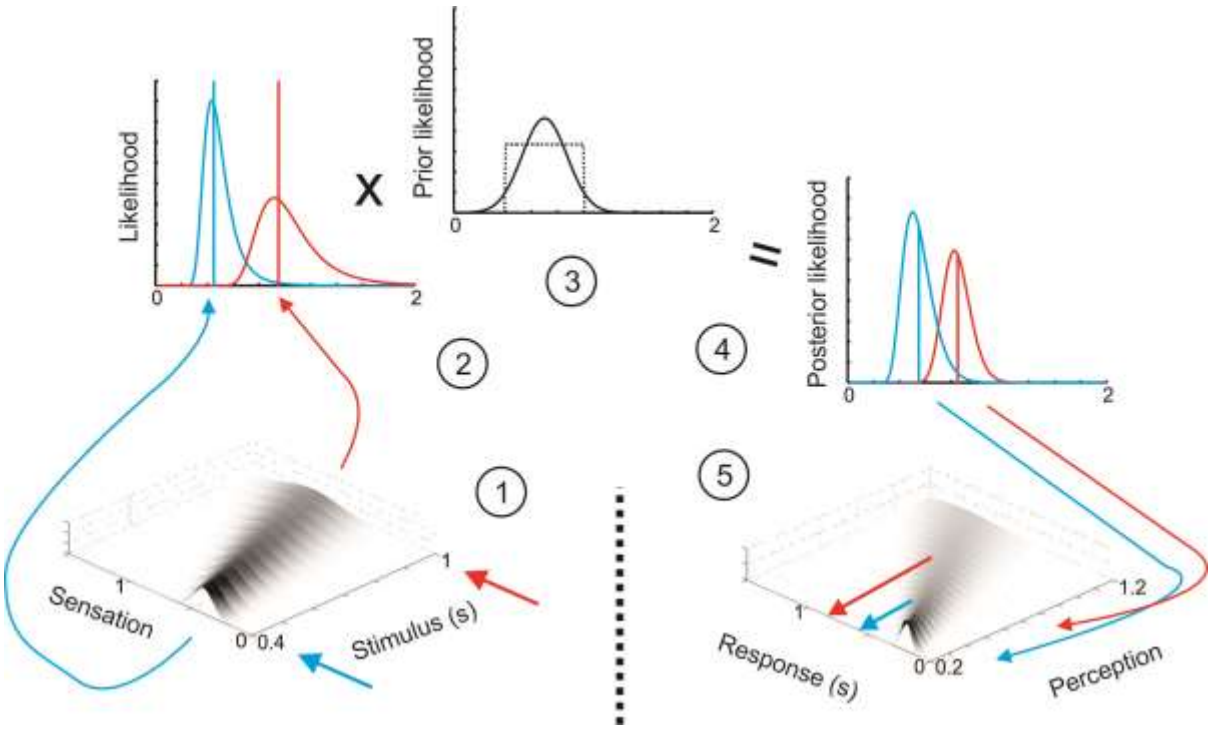

Figure 2. Schematic of a Bayesian observer model for duration reproduction (Jazayeri & Shadlen, 2010). Responses are traced through the model pipeline for two example stimuli, one of short duration (400 ms; blue) and one of long duration (1000 ms, red). Across multiple trials stimuli with these objective durations are corrupted by scalar sensory noise, to yield a range of sensory point estimates (with Gaussian distributions, as shown in the bottom-left plot, labelled 1). From each such point estimate, the Bayesian brain would use knowledge about the noise-generating process to infer likelihood functions (two examples are shown in the top-left plot, labelled 2). Likelihood functions are combined with priors: The top-middle plot (3) shows both the uniform distribution of stimulus durations used in the experiment, and the (assumed) Gaussian prior (which is derived from this uniform distribution). A point estimate is then recovered, based on the mean of the resulting posterior distribution (top-right plot; 4). Finally, these perceived durations are targeted, but further corrupted by scalar motor noise during reproduction, to yield the distribution of responses shown in the bottom-right plot (5).

The four-parameter model described so far can predict data from both auditory and visual blocks of the experiment, but only when the memory load is a single item. However, additional parameters, described below can capture behaviour with increased memory loads, and in intermixed blocks. Such high-parameter models are more challenging to fit. Hence, to better constrain fits from high-parameter models, the two sensory Weber ratios and the two motor Weber ratios outlined above were yoked by incorporating a punishment for a priori unlikely parameter combinations. Previous research indicates that for clearly supra-threshold stimuli, the precision of visual duration judgments

is somewhat worse than the precision of auditory duration judgments, with mean Weber ratios around 50% higher when estimated either with (e.g. Rammsayer, Buttkus, & Altenmüller, 2012) or without (e.g. Hartcher-O'Brien et al., 2014) trial-by-trial feedback. Hence we introduced an arbitrary penalty to the log-likelihood used during model fitting, based on the degree to which the ratio of visual to auditory sensory Weber ratios, r_{va}, deviated from the expected ratio, r_e = 1.5:

$$logL_{Penalised}(r_{va}|\mu,\sigma) = logL + \ln(f_r\left(\frac{r_{va}}{r_e};\ \mu = 1, \sigma = 1\right)^{10}) \quad (3)$$

Here f_r is the log-normal probability density function, which had parameters μ = 1 and σ = 1. A similar penalty was applied based on the two motor Weber ratios, but with a (presumed) large contribution from a common motor timer in both conditions, the predicted ratio was set as r_e = 1.0, i.e. the fitting procedure favoured motor noise estimates that were very similar between modalities. Note that none of our hypotheses are concerned with the exact values of participants' visual and auditory Weber ratios, nor with their ratio, so constraining these values should have little effect beyond preventing models from returning unlikely parameter values driven by noisy data.

2.6.3 Modelling (3): Performance with multiple items in memory

So far, we have described a model of behaviour in visual and auditory conditions with only a single item to reproduce. To extend the Bayesian-observer model to a working-memory task, we next considered the four memory architectures for which predictions are illustrated in Fig. 3. In memory slot models, memory stores are assumed to have a fixed number of available slots, which sets their capacity. In two variants of this account, the fifth and sixth model parameters represent slot number/capacity, *c*, for auditory and visual stimuli respectively. These were constrained to take on values between 0.9 (i.e. a capacity of 1, but with the possibility of some lapses) and 8 (the highest load, *l*, used in the current experiments, so no detectable capacity limit). On any given trial, the probability of having retained the target is:

$$p(t|l,c) = min\left(1, \frac{c}{l}\right) \quad (4)$$

Such that when load exceeds capacity, some items are forgotten. When the stimulus space is circular and sampled at random (i.e. the design typically applied in recent visual working memory studies; Ma et al., 2014), guessing will be uniform random. However, when the stimulus space is non-circular, as is the case for duration, the guessing strategy must be considered. We considered two (of many) possibilities. In the first, guessing occurred at random from the prior. In the second, guessing was always at the mean of the prior. Consequently, in a modelling step interposed between the derivation of a perceptual estimate from the posterior, and a noise-corrupted reproduction of this estimate, the predicted distribution of perceptual estimates, $p_{pe}(D_s)$, was mixed with the predicted distribution of guesses, $p_g(D_s)$, to yield the mixture density:

$$M(D_s) = p(t)p_{pe}(D_s) + (1 - p(t))p_g(D_s) \quad (5)$$

In this section, we have so far considered a slot model of working memory for duration. A different account of working memory posits *a resource that is shared* equally between all items, with this resource becoming less available in high-load conditions, which affects the precision with which items are represented in memory. To model the increase in memory noise in a compact fashion, we assumed a loaded sensory weber ratio increased with load as:

$$W_{s_l} = \frac{W_s}{a} \quad (6)$$

Where:

$$a = \left(\frac{1}{l}\right)^n \quad (7)$$

Here the exponent (*n*) controls the rate at which precision falls off. This relationship has previously proven plausible when modelling visual working memory (Bays & Husain, 2008). For two variants of

this memory-noise model, parameters 5 and 6 were the auditory and visual noise exponents respectively, constrained to take on values between 0 (i.e. no detectable increase in noise/decrease in precision) and 8 (an arbitrary upper limit representing a very rapid increase in noise). The two variants differed in terms of where memory noise was assumed to accrue, relative to the Bayesian estimation process. In a late Bayes integration variant, a sensory estimate already corrupted by memory noise was combined with a prior, as though integration occurred at the moment of recall. In an early Bayes integration variant, sensory variance was decomposed into a sensorial component, that accrued while the stimulus was encoded (prior to Bayes integration), and a memorial component (that accrued *after* Bayesian integration). To achieve this, the sensorial component (W_s) was used, even at loads above 1, to determine the likelihood function, and the remaining noise (estimated using the square root of W_{sl}^2- W_s^2) was applied after Bayes estimation.

The four models considered so far extend two prominent recent models of visual working memory to a task requiring reproduction of a non-circular stimulus attribute (i.e. duration). However, it is plausible that a slot limit might co-exist with decreasing memory resolution under increased load (Alvarez & Cavanagh, 2004). We therefore considered four further models, created by cross-combining models (i.e. each slot model with each memory noise model) to incorporate both a capacity limit and a noise exponent (for each modality). These more complex models had two further parameters (numbers 7 and 8; see Table 1).

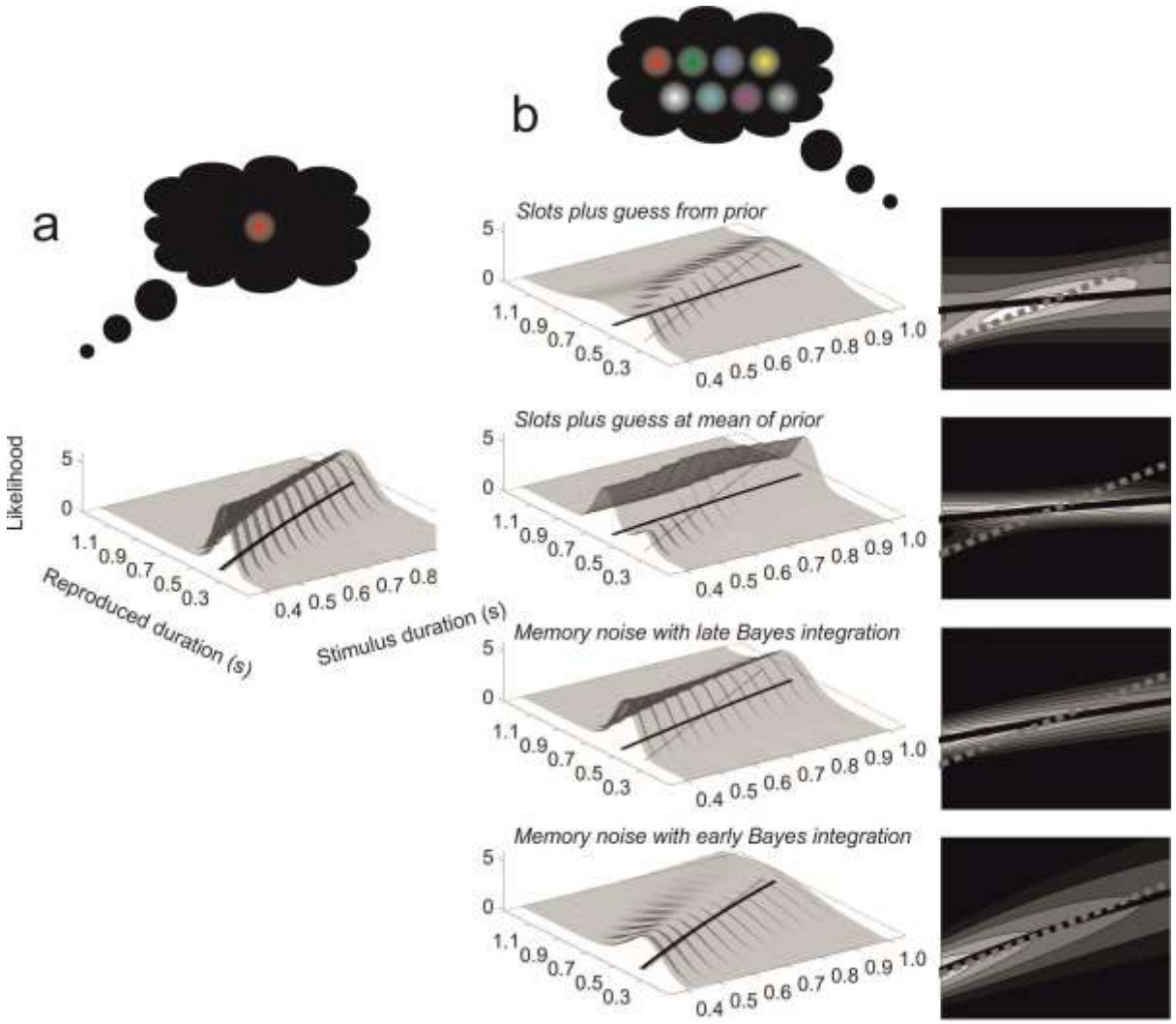

Figure 3. Illustrative predictions for four models of short-term memory for duration. Parameters used in simulations were: Sensory weber ratio = 0.1, motor weber ratio = 0.1, and either memory slot capacity = 2 or noise exponent = 0.5. Thin or dashed grey lines show objectively correct performance, i.e. reproduced duration = stimulus duration. Light grey 3D overlay or heat maps show predicted probability densities for responses made to stimuli with different durations. Thick black lines show the means of these predictions. ***A.*** *With only a single item in memory, all four models make an identical prediction, reflecting the Bayesian-observer model schematised in Figure 2. For a high-precision observer, the biasing effect of the prior is subtle (but slightly more evident at higher durations that accrue greater sensory noise) and mean performance is close to being objectively correct.* ***B.*** *When memory load is increased to eight items, performance is worse, and predictions vary for different models. From top to bottom: The slots plus guess from prior model shows highest probability density near the objectively correct value, but also a plateau region extending across the prior, which reflects guessing when memory capacity is exceeded (this dramatically affects the mean response); second plot down – the slots plus guess at mean of prior model shows greatest probability density near the mean of the prior (as with eight items in memory, and a two-item capacity, guessing is the most likely outcome) but also a marked probability of responding quite accurately (when the item has been remembered). This results in intermediate values for the mean response; third plot down – for the memory noise with late Bayes integration model, the increased noise in all memorised items yields a greater influence of the prior during the Bayesian integration process, and thus a shallower slope relating stimulus duration to mean reproduced duration; finally, fourth plot down – in the memory noise with early Bayes integration model, memory noise is added only after the*

Bayesian integration process, so the influence of the prior remains limited and mean performance shows little bias. There is, however, a variable error that increases dramatically relative to the single item case (note the shallower profile relative to the single case, depicted on the left).

2.6.4 Modelling (4): Performance with interleaved stimuli

Finally, *and of key importance for distinguishing our hypotheses,* all eight candidate models required a mechanism for incorporating trials from intermixed blocks and, critically, for assessing whether auditory and visual durations were being maintained in separate working-memory stores. To this end, we introduced a final model parameter (#7 for the four single-mechanism models, #9 for the four dual-mechanism models) – the intermixed load denominator, *d*. This parameter, applied only for intermixed trials, adjusted the load as follows:

$$l_{adjusted} = max\left(1, \frac{l}{d}\right) \quad (8)$$

The logic here is that if performance in intermixed trials declines with load in exactly the same manner as it does in unimodal trials, this behaviour will be captured with *d* = 1. On the other hand, if the intermixed condition allows participants to use two separate memory stores, performance will decline less precipitously, a behaviour that can be captured when *d* = 2. To provide room for random variation below and above these contrasting predictions, *d* was constrained to fall between 0 and 8.

3. Results

3.1 Preliminary analysis

Group averaged data are presented in Fig. 4. Fig. 4a shows our primary non model-based measure, which captures the magnitude of response errors across modality and memory-load conditions, for the main set of 30 participants. The intermixed condition has been further subdivided according to

whether participants had to reproduce an auditory or visual stimulus on a given trial. As anticipated, errors became more pronounced as memory load increased from 1 to 8 items, and were larger when reproducing visual compared to auditory stimuli. This modality difference was expected (e.g. Hartcher-O'Brien et al., 2014; Rammsayer, Buttkus, & Altenmüller, 2012) and is not in itself discriminatory for our hypothesis about memory stores, because performance will be worse if the encoded signal is less precise, even if the working memory mechanism is identical for visual and auditory intervals.

As illustrated in Fig. 4b, we also observed a tendency towards recency effects (i.e. better performance when the target appeared later in the stimulus sequence). Furthermore, reproductions showed a greater central tendency bias with increased load (Fig. 4c), no doubt generating some of the increase in root mean squared error already described (Fig. 4a).

Observations regarding how RMSE varied with modality and memory load (Fig. 4a) were confirmed via a 4x4 repeated-measures ANOVA (with Greenhouse-Geisser corrections for sphericity), which showed a main effect of load ($F_{[3,87]} = 19.06$, $p < 0.001$, partial $\eta^2 = 0.396$), a main effect of modality ($F_{[3,87]} = 20.84$, $p < 0.001$, partial $\eta^2 = 0.418$), but no interaction ($F_{[9,261]} = 1.036$, $p = 0.411$, partial $\eta^2 = 0.035$). Collapsing data across load, Tukey-corrected pairwise follow-up t-tests indicated that both auditory conditions differed from both visual conditions (all $p < 0.05$), but that being part of an intermixed (as opposed to a single-modality) block did not significantly affect errors for either auditory or visual stimuli (both $p > 0.18$). The failure to observe an interaction in this analysis suggests that intermixed blocks did not confer any advantage, in terms of mitigating the detrimental impact of a higher memory load, compared to blocks containing items from only a single modality (i.e. the slope of the solid and dashed lines in Fig. 4a is very similar).[3] This implies that participants

[3] The analysis of RMSE data can be reframed so that a two-store (as opposed to one-store) account provides the null hypothesis. To carry out such an analysis, we compared data from unimodal loads 1, 2 and 4, to that from intermixed loads 2, 4 and 8. The one-store account predicts a significant difference between such

may not have benefitted from an additional memory resource in these blocks. However, formal modelling was employed to better quantify the degree to which these data provide support for shared vs. independent memory resources for duration.

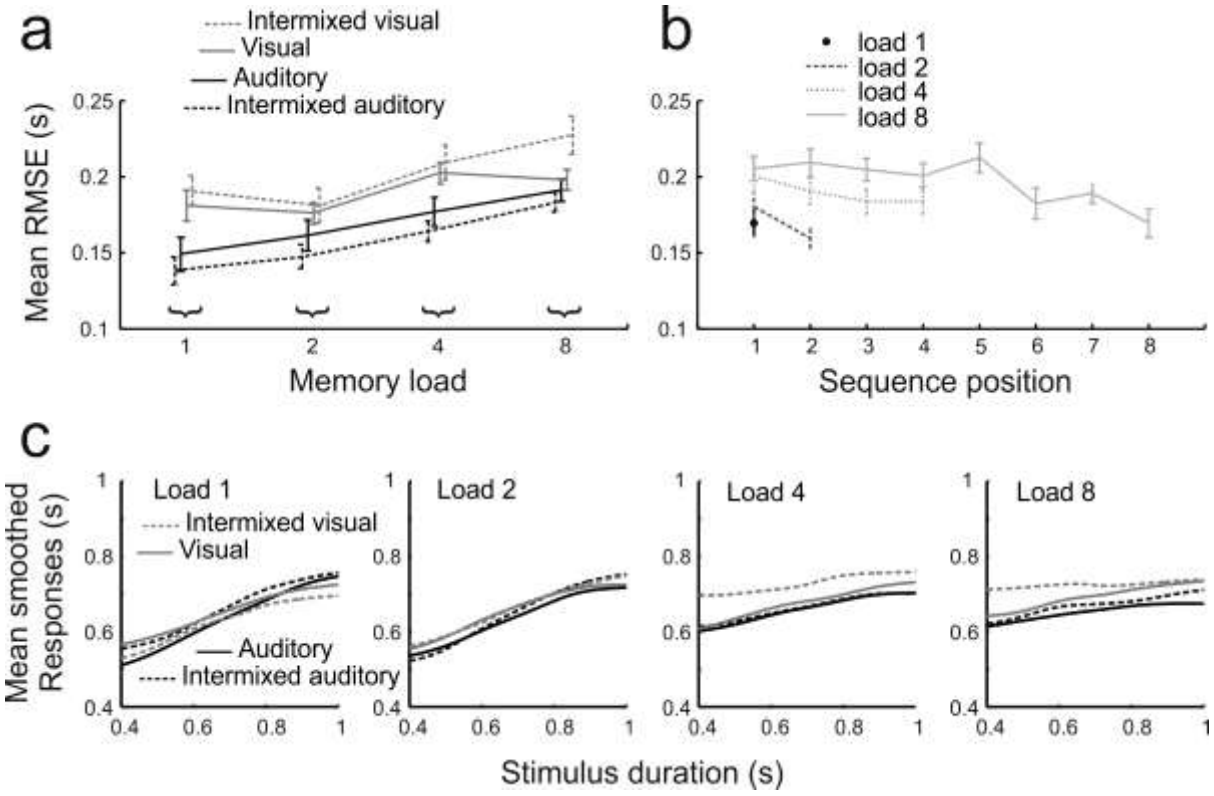

Figure 4. Group average data. Error bars denote standard error of the mean. ***A.*** *Average root mean squared errors (RMSE) of interval reproduction across 30 participants for auditory and visual stimuli, presented in either unimodal or intermixed blocks, and for memory loads of 1 to 8 items.* ***B.*** *Average RMSE of interval reproduction stratified by position of target within stimulus sequence and memory load (but collapsed across stimulus modality).* ***C.*** *Mean reproduced intervals in each condition. Prior to averaging across participants, responses on individual trials were smoothed into a moving average using a Gaussian kernel (cf. Kohl, Spieser, Forster, Bestmann, & Yarrow, 2019).*

3.2 Model selection

intermixed and unimodal conditions, because under this account, a lower load is being experienced in the unimodal case (whereas the two-store account, these conditions are effectively matched for load). A 3 (load) x 2 (block type: intermixed vs. unimodal) x 2 (modality) repeated-measures ANOVA revealed a main effect of block type ($F_{[1,29]} = 5.25$, $p = 0.029$) indicating that intermixed blocks (mean 0.186) generate higher RMSE scores than unimodal ones with half the load (mean 0.175). The ANOVA also revealed main effects of load ($F_{[2,58]} = 17.94$, $p < 0.001$) and modality ($F_{[1,29]} = 48.63$, $p < 0.001$) but no interactions (all $p > 0.13$).

Eight different models of working memory for duration were considered. The first four, with predictions illustrated in Fig. 3, represent either pure slot models (with different guessing strategies) or pure resource models (where Bayes integration is considered to occur either early, before encoding in memory, or late, when a memory is retrieved). The latter four models are more complex, and represent different possible combinations of the pure models, incorporating both a slot limit and a reduced precision with higher load. To assess which of these were viable, we calculated group-mean AIC and BIC values, from the maximum-likelihood fits to each participant's data. Lower values indicate better fits. These metrics, presented in Fig. 5, implement different theoretically derived corrections for the number of parameters in order to equate model complexity, with BIC punishing additional parameters more aggressively than AIC. While AIC and BIC suggest different winning models, a clear pattern emerges across both metrics – both for the two observers who completed multiple blocks of trials, and for the group of 30 participants who completed a single block per condition, three models emerge as weak performers relative to the other five.

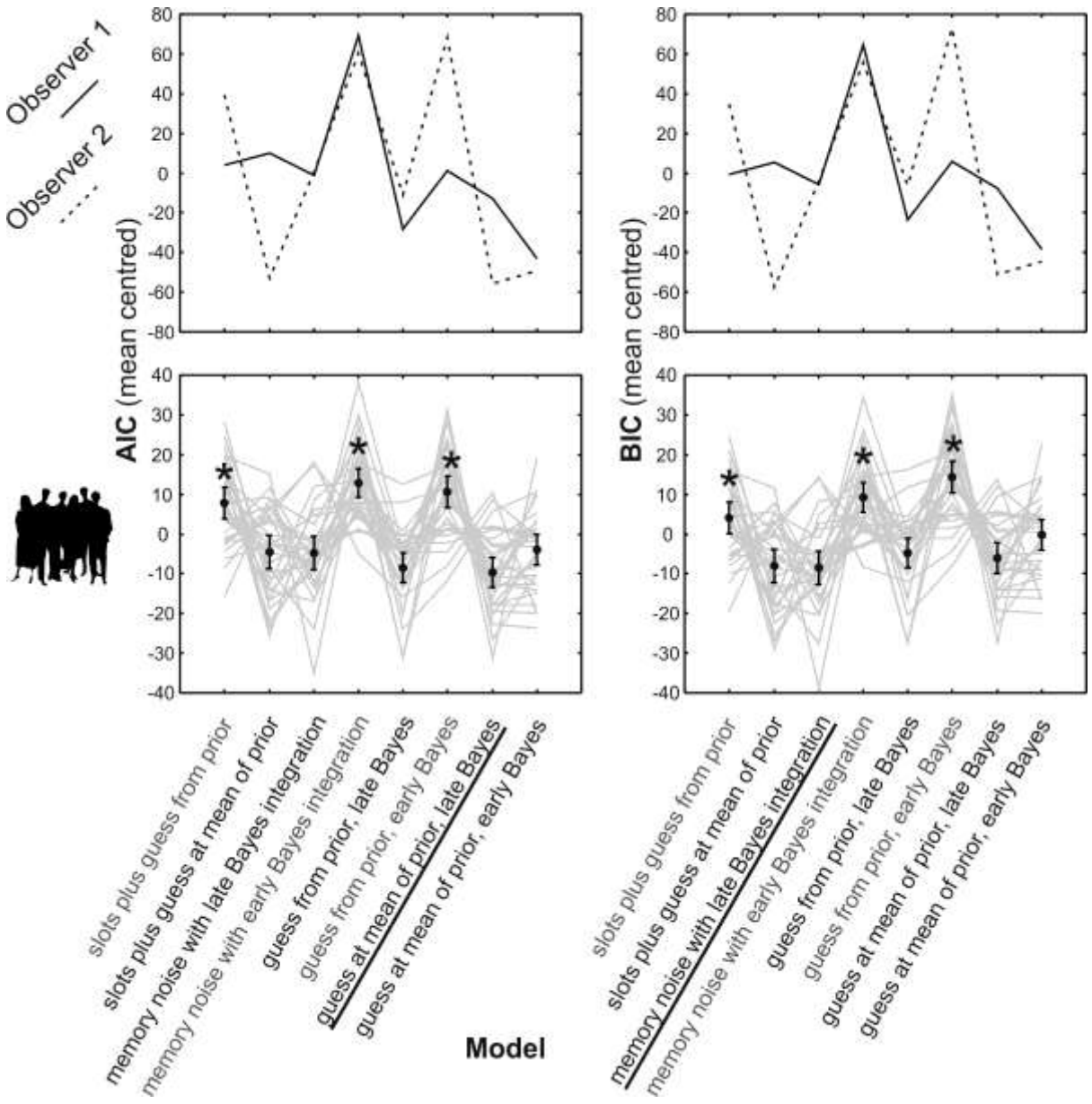

Figure 5. Mean-centred AIC (left) and BIC (right) values for best-fitting variants of eight models of short-term memory for duration. Models 1-4 implement either slot limits or memory noise to explain performance decrements with increasing load, whereas models 5-8 cross combine these two kinds of modelling approach. The mean value across models for each participant has been subtracted from their scores to increase y-axis resolution. Upper plots show data for the two observers who completed additional blocks for illustrative purposes. Lower plots show data for a further 30 participants (grey lines), along with their group average scores (black lines) and 95% confidence intervals (error bars). Winning models at the group level are underlined. Models with significantly higher AIC/BIC values than these winning models (Tukey's honestly significant $p < 0.05$) are denoted by grey text and asterisks (*).

Group-level data were subjected to repeated-measures ANOVAs, which indicated significant Greenhouse-Geisser corrected differences between models on both metrics (AIC: $F_{[7,203]} = 19.28$, $p < 0.001$; BIC: $F_{[7,203]} = 17.22$, $p < 0.001$). Pairwise follow-up Tukey tests indicated that, based on AIC,

five models, including two simpler models (*slots plus guess at mean of prior; memory noise with late Bayes integration*) and three more complex models (*guessing from prior combined with late Bayes integration; guessing at mean of prior combined with early Bayes integration; guessing at mean of prior combined with late Bayes integration*) did not differ from each other, but were all significantly better supported than the remaining three models (all $p < 0.05$). The pattern for BIC was a little more complex, but the same three weaker models emerged (*slots plus guess from prior; memory noise with early Bayes integration; guessing from prior combined with early Bayes integration*), which could all be rejected ($p < 0.05$) relative to the best fitting model (*memory noise with late Bayes integration*). Two of the weaker models could also be rejected relative to three of the other four plausible models, and one could be rejected relative to the final plausible model (*guessing at mean of prior combined with early Bayes integration*).

3.3 Model illustration

Although these analyses suggest that five models are plausible, to remain succinct we present illustrative data from only the *guess at mean of prior combined with late Bayes integration* model, which won at the group level based on AIC (and was also a good performer for the two observers who completed multiple blocks of trials). These illustrative data are shown in Figs. 6 – for auditory conditions, and 7 – for intermixed conditions with auditory targets. These two figures together depict half of the data collected for each of these observers (visual target conditions are presented in Appendix A for completeness), and they illustrate a good overlap between model predictions and data.

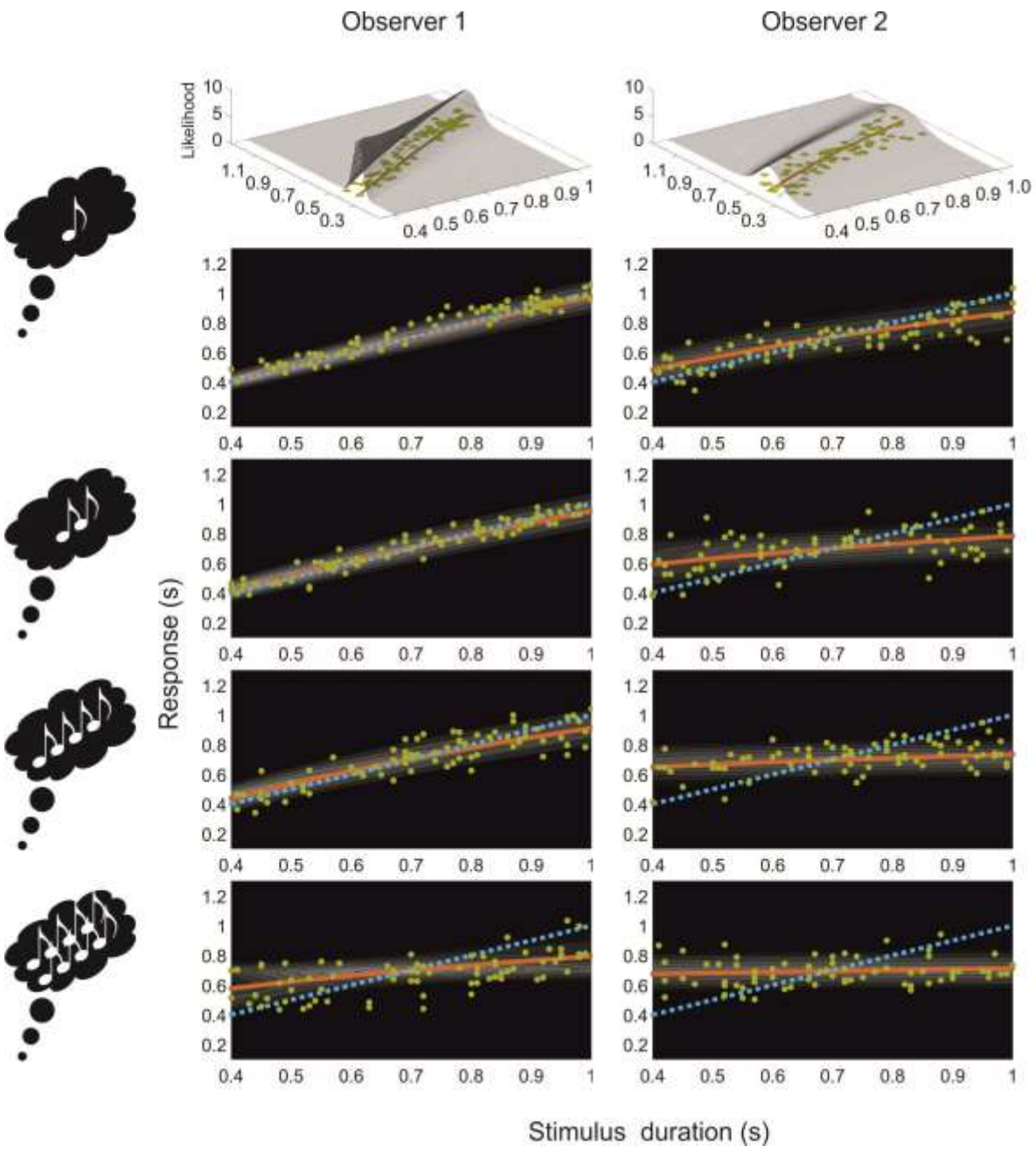

Figure 6. Model predictions and data from auditory conditions for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both of these observers. This model also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 2, for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Best-fitting auditory parameters for observer 1: Sensory Weber ratio 0.053; motor Weber ratio 0.070; noise exponent 0.640; slot capacity 4.827. Best-fitting auditory parameters for observer 2: Sensory Weber ratio 0.181; motor Weber ratio 0.120; noise exponent 0.174; slot capacity 1.084.

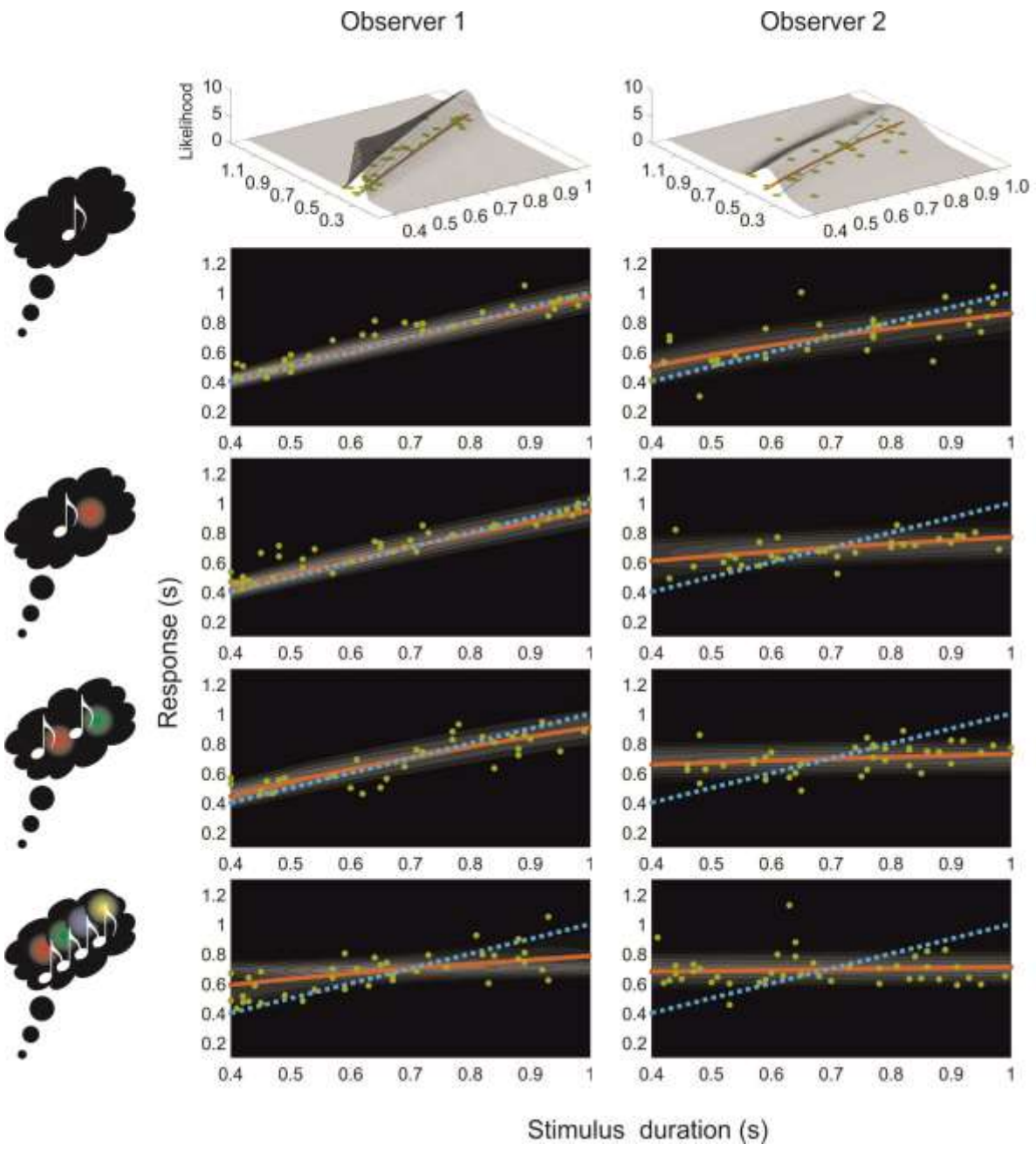

Figure 7. Model predictions and data from auditory judgements made in intermixed conditions, for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both observers, and which also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 2 for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Model predictions vary very slightly from those shown in Figure 6 because load is affected by the intermixed load denominator parameter (observer 1: 0.934; observer 2: 0.862).

3.4 Model parameters and implications

Median best-fitting parameters (from the model illustrated in Figs. 6 & 7) for the group as a whole are presented in Fig. 8a.[4] In general, parameter estimates seem sensible. For example, Weber ratios, which capture sensory precision, took on values of ~0.2. Although these are not directly comparable to estimates from other kinds of task/analyses, these are broadly in line with previous research using untrained participants on different tasks (e.g. Ball et al., 2017; Narkiewicz, Lambrechts, Eichelbaum, & Yarrow, 2015), and they are a little above the more exactly equivalent estimates from work adopting a similar approach (but with more practised participants, with no memory manipulation; Acerbi et al., 2012). A parameter recovery simulation (Appendix B) indicated that Weber ratios could be recovered successfully using our methods.

Several parameters suggest some degree of bimodality across the group, including slot capacities, which determine when participants must resort to guessing, and noise exponents, which control the rate at which precision declines with increasing load. We should be cautious in interpreting these parameters, which could be recovered less well (Appendix B) and only with a degree of artefactual clustering at the extremes. However, the spread and bimodality may also reflect tendencies for the behavioural decline of different participants to be captured mainly by either a slot capacity limit (in which case the noise exponent will be near zero, where it does not affect behaviour), or by decreasing precision (in which case the slot capacity will be near 8, where it does not affect behaviour). These conclusions are supported by the positive correlation between these parameters across participants for both auditory ($r = 0.58$) and visual ($r = 0.78$) trials.

[4] Medians were preferred to means because distributions were heavily skewed, particularly for the critical intermixed load denominator parameter which we use for statistical inference.

Of particular interest is the value of the final model parameter, the intermixed load denominator. When this is above 1, it reduces the effective load experienced in intermixed conditions. A value of 2.0 is expected if intermixed stimuli from each modality are maintained completely independently in two separate stores, and a value of 1.0 is predicted if they are maintained in a single store. Simulations revealed that this parameter was well recovered at the group level (Appendix B). We can therefore assess whether either account is inconsistent with data at the 5% level by determining 95% confidence intervals about this parameter (here achieved using bias-corrected and accelerated bootstrapping). For the *guess at mean of prior combined with late Bayes integration* model, we can reject the independent-store account.

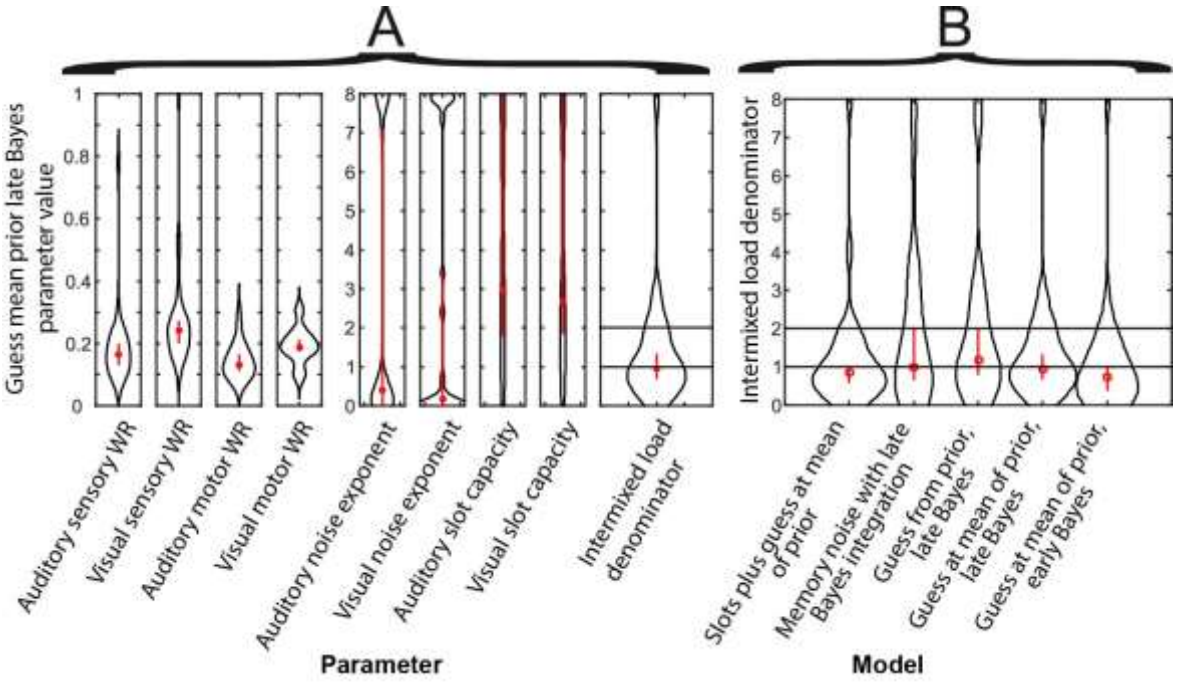

Figure 8. Kernel density plots summarising group parameter values from best-fitting models. The widths of black outline shapes represent probability densities. Red markers indicate medians. Red vertical bars indicate 95% bootstrap confidence intervals around medians. For intermixed load denominator parameter plots, dashed horizontal lines at values of 1 and 2 indicate predictions for single-store and dual-store accounts, respectively. WR = weber ratio. ***A.*** *All nine parameters are illustrated for the well-supported guess at mean of prior, late Bayes integration model. The intermixed load denominator parameter does not differ significantly from 1.0, but does differ significantly from 2.0, supporting a single-store account.* ***B.*** *The intermixed load denominator parameter for all five plausible models (see Figure 4). Confidence intervals rarely overlap 2, but almost always overlap 1, again supporting a single-store account.*

The model illustrated in Figs. 6 and 7 was not the only plausible model for our data. Hence Fig. 8 also shows the intermixed load denominator parameter, which tests the single and dual-store accounts, for each of the five plausible models. The trend across all five models is clear, although occasionally marginal. The independent-store account can generally be rejected, with its prediction falling outside (or occasionally at the extreme margin of) the 95% confidence interval. By contrast, the median value for this parameter is generally centred quite near the prediction of the single-store account (i.e. d = 1.0). This group-level conclusion was also supported by analyses of data for our two observers, for whom the mean denominator across the five plausible models was 0.998 (SD 0.089 – observer 1) and 0.799 (SD 0.174 – observer 2).

3.5 Positive control experiment

Although our parameter recovery simulation (Appendix B) suggested that the critical d model parameter should have been recoverable (showing a value close to 2.0) if participants had shown improvement consistent with the existence of two stores in the intermixed condition, we also sought an empirical demonstration. This is important because, as we expand in the discussion, the expectation that two memory stores would effectively half the load may be overly optimistic. For example, in intermixed conditions the sequence will have a longer duration than the equivalent half-length sequence from unimodal conditions, so even if load is equated (via the presence of two stores) retention time will not be, perhaps depressing performance. To address this issue, we ran a small control study in which a fourth condition was introduced, involving intermixed stimuli, but with an accurate pre-cue alerting participants to which stimulus modality would contain the target. This should have allowed them to strategically half their load (by attending to just one modality) so long as the increase in overall sequence duration did not mask any such effect by substantially increasing overall difficulty.

For the five plausible models identified from our main experiment, participants in the control experiment returned a mean *d* of 1.67 (95% CI 0.84-2.49) in the new cued intermixed condition, and a mean *d* of 0.93 (95% CI 0.61-1.26) in the original intermixed condition. They were in general best fitted by the memory noise with late Bayes integration model (best for 3/4 by BIC, 2/4 by AIC) where the corresponding *d* estimates were 1.88 (95% CI 1.16-2.60) and 0.88 (95% CI 0.74-1.03) respectively. These results suggest that our prediction (for a *d* of 2.0 if stimuli could be sorted into two stores) was slightly overoptimistic, given task challenges (e.g. the effective change in sequence duration). However, as can be seen in Fig. 8b, setting the expectation for a two-store *d* to a lower level (e.g. 1.67) would still have implied rejection of this account under three out of five plausible models.

4. Discussion

Here we have examined working memory for the duration of inputs, presented solely in auditory or visual modalities, or split between these modalities. We have found that reproduction performance gets worse as memory load is increased, but to a similar extent when items could in principle be transferred into two separate modality-specific stores (should they exist). We then applied formal modelling, to identify plausible memory architectures for duration, and to exclude implausible architectures. Several models provided a plausible account of data, and these could not be discriminated from one another. These included a slot model, a resource model, and three models that combined both of these processes. However, across models, analyses tended to support the idea that the durations of stimuli from both modalities were maintained in a single store, that drew upon a common memory resource.

The finding that both categories of input appear to be encoded in a common store is at odds with work by Rattat and Picard (2012), who found selective interference with auditory interval

comparison by an articulatory suppression task, and with visual interval comparison by a visuospatial tracking task. However, their result could not be replicated by Bratzke et al. (2016), who used a more powerful repeated-measures design and instead found an effect of articulatory suppression (but not visuospatial tracking) for both modalities. Our result is most consistent with Bratzke et al.'s (2016) finding, and taken together these findings suggest the possibility that short-term memory for duration is not strictly amodal, but rather may sometimes involve a translation of non-auditory stimuli into an auditory format for storage/manipulation. However, a subsequent study (Bratzke & Ulrich, 2020) with different secondary tasks found that pitch discrimination affected only auditory interval comparison, with colour discrimination not affecting interval comparison in either modality, and auditory interval comparison negatively affecting RT on both auditory and visual secondary tasks. Hence further studies are clearly warranted to help pick apart conflicting results. Other work has suggested an amodal (or shared) storage of durations over longer timescales, for example the proactive influence of auditory stimuli in one block on visual stimuli presented in the next block, and vice versa (Filippopoulos et al., 2013). Their result might imply that the prior for duration is generally updated without regard to modality (see also Roach, McGraw, Whitaker, & Heron, 2017).

Determining whether the durations of stimuli from different modalities enter distinct or common memory stores helps address an important question in the timing literature: Whether timed behaviours depend on a single internal clock, or arise from a variety of (perhaps co-ordinated) modality-specific timing mechanisms (Ivry & Schlerf, 2008). Previous work (e.g. Ball et al., 2017; Hartcher-O'Brien et al., 2014) has suggested that information about duration can be combined from each of the unimodal signals comprising a bimodal stimulus, improving precision. Hence the noise that limits performance must, at some level, be independent for each modality. The current results suggest that the working memory components are not independent, and thus, in combination with previous findings, imply that the underlying pacemakers may constitute a modality specific

component of the interval-timing system. This would be consistent with intrinsic models of time perception (e.g. Buonomano & Merzenich, 1995; Roseboom et al., 2019).

Previous studies have examined working memory for the duration of multiple items using a time reproduction task, but only in the auditory modality, and without applying formal modelling to draw inferences about plausible memory architectures (Manohar & Husain, 2016; Teki & Griffiths, 2014; Teki & Griffiths, 2016). We addressed this oversight, by adopting a formal modelling approach, as model predictions can be hard to intuit for all but the simplest of situations. A non-circular stimulus space makes predictions for different classes of model more similar to each other, relative to circular attributes like colour and orientation (commonly assessed in recent working-memory literature; see Ma et al., 2014). Given the lack of consensus that exists in that literature, it is perhaps not surprising that we were unable to find a clear winning model here. In fact, both slot models and resource models proved viable. We were, however, able to reject three models, and these rejections provide some interesting insights. For example, a pure resource model was only viable if the likelihood function representing the sensed duration was combined with a prior belatedly, at the point at which the stimulus was retrieved from memory. We suspect this is rather later than most proponents of Bayesian-observer models would envisage (e.g. Weiss, Simoncelli, & Adelson, 2002) but of course these descriptive models tend to be strictly agnostic on this point. A more standard account (i.e. early integration) probably failed because it did not capture the increasing central tendency bias observed with increasing memory load (compare Figure 3b, bottom illustration, with results in Figure 4c), whereas our other models did.

The conclusions we have reached are limited by a variety of factors. First, we can consider limits to the modelling. The front end of our model, the Bayesian observer, assumed both a Gaussian approximation of the prior and scalar (i.e. multiplicative) timing noise (because both assumptions had found previous support; Acerbi et al., 2012; Cicchini et al., 2012; Jazayeri & Shadlen, 2010). At

heart, this model assumes a prior based on an immediate and accurate knowledge of the current temporal context, which of course must be a simplification of the way priors are established and updated (cf. Bausenhart, Dyjas, & Ulrich, 2014; Di Luca & Rhodes, 2016). However, this is not unreasonable given that participants both received practice and saw large numbers of stimuli quite quickly as a result of high-load trials. Nonetheless, other choices for modelling the Bayes observer (or duration reproduction in general) might have encouraged different conclusions.

We also note that the model does not incorporate parameters to capture over or under-reproduction biases operating uniformly across the full range of stimuli (because feedback is assumed to eliminate such biases). Given that sounds are in general judged longer than lights of equivalent duration, a relative bias might emerge between them, particularly in intermixed blocks, and could be worth modelling in future developments of this work. However, while this is certainly a valid concern, we do not think this oversight is likely to have led to an erroneous conclusion here, for two reasons. First, in response to a reviewer's query we determined the mean relative bias in reproduction errors between auditory and visual stimuli in our intermixed blocks. This relative bias was small, failing to reach statistical significance for our sample of 30 participants, perhaps because feedback was fairly effective in eliminating it. Second, in order to falsely suppress the load denominator parameter that we recovered and used to test our primary hypothesis, any source of overlooked noise in the modelling of our intermixed conditions would need to be more noticeable at higher loads. It seems unlikely that a relative modality bias would have this characteristic.

Similarly, while we considered some prominent models of working memory, and their combination, our eight models inevitably leave myriad possibilities untested. For example, we did not model transposition/binding errors, i.e. confusions of the target with another stimulus (Bays, Catalao, & Husain, 2009). Neither did we consider more complex accounts of how resources might be shared, for example the idea that sharing is unequal, leading to a mixture-of-Gaussian prediction for a

circular stimulus space (Fougnie, Suchow, & Alvarez, 2012; van den Berg, Shin, Chou, George, & Ma, 2012). Both ideas are likely relevant, particularly given that both primacy and recency effects have been found in experimental protocols similar to ours (Fan & Yotsumoto, 2018; Manohar & Husain, 2016; Teki & Griffiths, 2014; Teki & Griffiths, 2016). These order effects (or at least recency) can also be seen in Fig. 4c here, and should ideally be modelled to better capture this source of noise. Furthermore, we have not attempted to formally model any categorisation of duration at discrete attraction points (see comments on possible verbal recoding, below). This has proved valuable when modelling other continuous stimulus attributes such as colour (e.g. Hardman, Vergauwe, & Ricker, 2017) although the tendency to categorise may be more natural for colour than for duration, given the existence of discrete linguistic labels with agreed absolute meanings. It may not be possible to adjudicate slot from resource models without modelling some of these additional features.

In addition to the modelling, we must also consider limitations to our experimental approach. An obvious objection is that one might expect some negative impact of intermixed stimuli, even if they are encoded in separate, independent stores. There might be a cost to switching attention back and forth between distinct modes of input. Furthermore, sequential presentation, which is the only option if duration is to be encoded successfully (Morgan, Giora, & Solomon, 2008), dictates that intermixed load-eight conditions are longer than unimodal load-four conditions, implying a longer retention period, which might yield a greater accumulation of internal noise (cf. Gamache & Grondin, 2010). Such differences between the conditions we compared might also be exacerbated by the modelling limitations we have already discussed. For example, it may be that (unmodelled) transposition errors would be both more prevalent for higher loads (which potentially increase item position confusion) and more damaging in intermixed conditions (where colour/tone cues might be more difficult to utilise, and the adjustment after seeing a reinstated stimulus from an unexpected modality might be more jarring). Such factors could counteract the benefits of separate stores (if present).

These are clearly valid concerns (and might even explain why estimates of the intermixed load-division parameter was sometimes < 1.0, implying participants got worse in intermixed conditions). However, we went some way towards addressing some of them through the inclusion of our positive control experiment, which recovered a division parameter well above 1.0 (and nearer to 2.0) when participants were pre-cued about the modality of the target in intermixed blocks. It hence seems fairly unlikely that these factors would have effectively doubled load (under a scenario of two independent stores), and thus created the false impression of there being just one amodal store.

A further concern relates to the possible recoding of stimuli in our experiment. It is possible, for example, that without a secondary task (such as articulatory suppression), participants converted duration estimates into a verbal code, using their phonological loop. The phonological loop would then become *the* store for duration (and hence unitary, but not amodal). For this to happen, a profound form of translation would have to have been implemented, with both kinds of input converted into a verbal label expressing a relative duration (e.g. “short-long-medium-long”). Our data do not speak for, or against this possibility. Like Manohar and Husain (2016), but unlike Teki and Griffiths (2014; 2016), we included unfilled breaks between filled stimuli. This has the disadvantage of allowing greater time for a recoding of stimuli, but has the advantage of making it less likely that stimuli can be chunked into a higher-order stimulus train, akin to a rhythm.

A final methodological issue concerns the roles of instruction and cuing/binding in our experiment. We aimed to maximise performance, and reasoned that we could do so by having both stimulus position, and either colour or pitch, act as memory cues for recall. One might think that providing an instruction to participants regarding the upcoming sequence length might further improve performance, but a (hard to explain) result from Manohar and Husain (2016) suggests otherwise. These authors found that providing an instruction about sequence length *reduced* performance,

specifically in the load-one condition, which might lead to an underestimate of the effect of memory load (something we attempted to avoid).

In summary, we have combined a Bayesian observer model with several candidate accounts of working memory for duration. We identified several plausible memory architectures for this little-researched stimulus attribute. Furthermore, by exposing participants to sequences of either auditory, visual, or intermixed stimuli, and having them attempt to reproduce one of these at random, we have provided support for the idea that durations, encoded from different modalities, are stored in a single limited-capacity working-memory store. This conclusion is supported by both model and non-model based analyses of our data. The precise manner in which durations are retained requires further elucidation.

Acknowledgements

The authors would like to thank Ansgar Endress and Joshua Solomon for their constructive comments during the drafting process.

5. References

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology, 8*(11), e1002771.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106-111. doi:10.1111/j.0963-7214.2004.01502006.x

Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research, 150*(0079-6123), 45-53.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology, 63*, 1-29.

Ball, D. M., Arnold, D. H., & Yarrow, K. (2017). Weighted integration suggests that visual and tactile signals provide independent estimates about duration. *Journal of Experimental Psychology. Human Perception and Performance, 43*(5), 868-880.

Bartolo, R., & Merchant, H. (2009). Learning and generalization of time production in humans: Rules of transfer across modalities and interval durations. *Experimental Brain Research, 197*(1), 91-100.

Bausenhart, K. M., Dyjas, O., & Ulrich, R. (2014). Temporal reproductions are influenced by an internal reference: Explaining the vierordt effect. *Acta Psychologica, 147*, 60-67.

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 11.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science, 321*(5890), 851-854.

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

Bratzke, D., Quinn, K. R., Ulrich, R., & Bausenhart, K. M. (2016). Representations of temporal information in short-term memory: Are they modality-specific? *Acta Psychologica, 170*, 163-167.

Bratzke, D., Seifried, T., & Ulrich, R. (2012). Perceptual learning in temporal discrimination: Asymmetric cross-modal transfer from audition to vision. *Experimental Brain Research, 221*(2), 205-210.

Bratzke, D., & Ulrich, R. (2020). Short-term memory of temporal information revisited. *Psychological Research*. https://doi.org/10.1007/s00426-020-01343-y

Bryce, D., Seifried-Dübon, T., & Bratzke, D. (2015). How are overlapping time intervals perceived? Evidence for a weighted sum of segments model. *Acta Psychologica, 156*, 83-95.

Buonomano, D. V., & Merzenich, M. M. (1995). Temporal information transformed into a spatial code by a neural network with realistic properties. *Science, 267*(5200), 1028-1030.

Cicchini, G. M., Arrighi, R., Cecchetti, L., Giusti, M., & Burr, D. C. (2012). Optimal encoding of interval timing in expert percussionists. *The Journal of Neuroscience, 32*(3), 1056-1060. doi:10.1523/JNEUROSCI.3411-11.2012

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences, 24*(1), 8-185.

Cowan, N., Saults, J. S., & Blume, C. L. (2014). Central and peripheral components of working memory storage. *Journal of Experimental Psychology. General, 143*(5), 1806-1836. doi:10.1037/a0036814

Di Luca, M., & Rhodes, D. (2016). Optimal perceived timing: Integrating sensory information with dynamically updated expectations. *Scientific Reports, 6*, 28563.

Endress, A. D., Korjoukov, I., & Bonatti, L. L. (2017). Category-based grouping in working memory and multiple object tracking. *Visual Cognition, 25*(9-10), 868-887. doi:10.1080/13506285.2017.1349229

Fan, Z., & Yotsumoto, Y. (2018). Multiple time intervals of visual events are represented as discrete items in working memory. *Frontiers in Psychology, 9*, 1340.

Filippopoulos, P. C., Hallworth, P., Lee, S., & Wearden, J. H. (2013). Interference between auditory and visual duration judgements suggests a common code for time. *Psychological Research, 77*(6), 708-715.

Fougnie, D., Suchow, J. W., & Alvarez, G. A. (2012). Variability in the quality of visual working memory. *Nature Communications, 3*, 1229.

Fougnie, D., Zughni, S., Godwin, D., & Marois, R. (2015). Working memory storage is intrinsically domain specific. *Journal of Experimental Psychology: General, 144*(1), 30-47.

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review, 17*(5), 673-679.

Gamache, P., & Grondin, S. (2010). Sensory-specific clock components and memory mechanisms: Investigation with parallel timing. *European Journal of Neuroscience, 31*(10), 1908-1914.

Gibbon, J., Church, R. M., & Meck, W. H. (1984). Scalar timing in memory. *Annals of the New York Academy of Sciences, 423*, 52-77.

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance*, 43(1), 30.

Hartcher-O'Brien, J., Di Luca, M., & Ernst, M. O. (2014). The duration of uncertain times: Audiovisual information about intervals is integrated in a statistically optimal fashion. *PloS One, 9*(3), e89339.

Heron, J., Aaen-Stockdale, C., Hotchkiss, J., Roach, N. W., McGraw, P. V., & Whitaker, D. (2012). Duration channels mediate human time perception. *Proceedings of the Royal Society B: Biological Sciences, 279*(1729), 690-698. doi:10.1098/rspb.2011.1131

Ivry, R. B., & Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences, 12*(7), 273-280.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience, 13*(8), 1020-1026.

Kohl, C., Spieser, L., Forster, B., Bestmann, S., & Yarrow, K. (2019). The neurodynamic decision variable in human multi-alternative perceptual choice. *Journal of Cognitive Neuroscience, 31*(2), 262-277.

Lapid, E., Ulrich, R., & Rammsayer, T. (2009). Perceptual learning in auditory temporal discrimination: No evidence for a cross-modal transfer to the visual modality. *Psychonomic Bulletin & Review, 16*(2), 382-389.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279-281.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347-356.

Manohar, S. G., & Husain, M. (2016). Working memory for sequences of temporal durations reveals a volatile single-item store. *Frontiers in Psychology, 7*, 1655.

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*(2), 81-97.

Morgan, M. J., Giora, E., & Solomon, J. A. (2008). A single "stopwatch" for duration estimation, a single "ruler" for size. *Journal of Vision, 8*(2), 14-18.

Nagarajan, S. S., Blake, D. T., Wright, B. A., Byl, N., & Merzenich, M. M. (1998). Practice-related improvements in somatosensory interval discrimination are temporally specific but generalize across skin location, hemisphere, and modality. *The Journal of Neuroscience, 18*(4), 1559-1570.

Narkiewicz, M., Lambrechts, A., Eichelbaum, F., & Yarrow, K. (2015). Humans don't time subsecond intervals like a stopwatch. *Journal of Experimental Psychology: Human Perception and Performance, 41*(1), 249.

Ogden, R. S., Wearden, J. H., & Jones, L. A. (2010). Are memories for duration modality specific? *Quarterly Journal of Experimental Psychology, 63*(1), 65-80.

Palmer, J. (1990). Attentional limits on the perception and memory of visual information. *Journal of Experimental Psychology. Human Perception and Performance, 16*(2), 332-350.

Paton, J. J., & Buonomano, D. V. (2018). The neural basis of timing: distributed mechanisms for diverse functions. *Neuron, 98*(4), 687-705.

Price, K. V., Storn, R. N., & Lampinen, J. A. (2006). *Differential evolution: A practical approach to global optimization*. Heidelberg: Springer.

Rammsayer, T. H., Buttkus, F., & Altenmüller, E. (2012). Musicians do better than nonmusicians in both auditory and visual timing tasks. *Music Perception: An Interdisciplinary Journal, 30*(1), 85-96.

Rattat, A., & Picard, D. (2012). Short-term memory for auditory and visual durations: Evidence for selective interference effects. *Psychological Research, 76*(1), 32-40.

Roach, N. W., McGraw, P. V., Whitaker, D. J., & Heron, J. (2017). Generalization of prior information for rapid bayesian time estimation. *Proceedings of the National Academy of Sciences, 114*(2), 412-417.

Roseboom, W., Fountas, Z., Nikiforou, K., Bhowmik, D., Shanahan, M., & Seth, A. K. (2019). Activity in perceptual classification networks as a basis for human subjective time perception. *Nature Communications, 10*(1), 267.

Teki, S., & Griffiths, T. D. (2014). Working memory for time intervals in auditory rhythmic sequences. *Frontiers in Psychology, 5*, 1329.

Teki, S., & Griffiths, T. D. (2016). Brain bases of working memory for time intervals in rhythmic sequences. *Frontiers in Neuroscience, 10*, 239.

Uittenhove, K., Chaabi, L., Camos, V., & Barrouillet, P. (2019). Is working memory storage intrinsically domain-specific? *Journal of Experimental Psychology: General (epub ahead of print)*

van den Berg, R., Shin, H., Chou, W., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences of the United States of America, 109*(22), 8780-8785.

van Rijn, H., & Taatgen, N. A. (2008). Timing of multiple overlapping intervals: How many clocks do we have? *Acta Psychologica, 129*(3), 365-375.

Wearden, J. H., & Lejeune, H. (2008). Scalar properties in human timing: Conformity and violations. *Quarterly Journal of Experimental Psychology, 61*(4), 569-587.

Weiss, Y., Simoncelli, E. P., & Adelson, E. H. (2002). Motion illusions as optimal percepts. *Nature Neuroscience, 5*(6), 598-604.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 1120-1135.

Wong, J. H., Peterson, M. S., & Thompson, J. C. (2008). Visual working memory capacity for objects from different categories: A face-specific maintenance effect. *Cognition, 108*(3), 719-731.

Wood, J. N. (2007). Visual working memory for observed actions. *Journal of Experimental Psychology. General, 136*(4), 639-652.

Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*(7192), 233-235.

Appendix A

Illustrative data from the *guess at mean of prior combined with late Bayes integration* model, which won at the group level based on AIC (and was also a good performer for the two observers who completed multiple blocks of trials). These illustrative data are shown in Figures A1 – for visual conditions, and A2 – for intermixed conditions with visual targets. These two figures together depict the remaining half of the data collected for each of these observers (data and fits from auditory target conditions are illustrated in the main text, Figures 6 and 7).

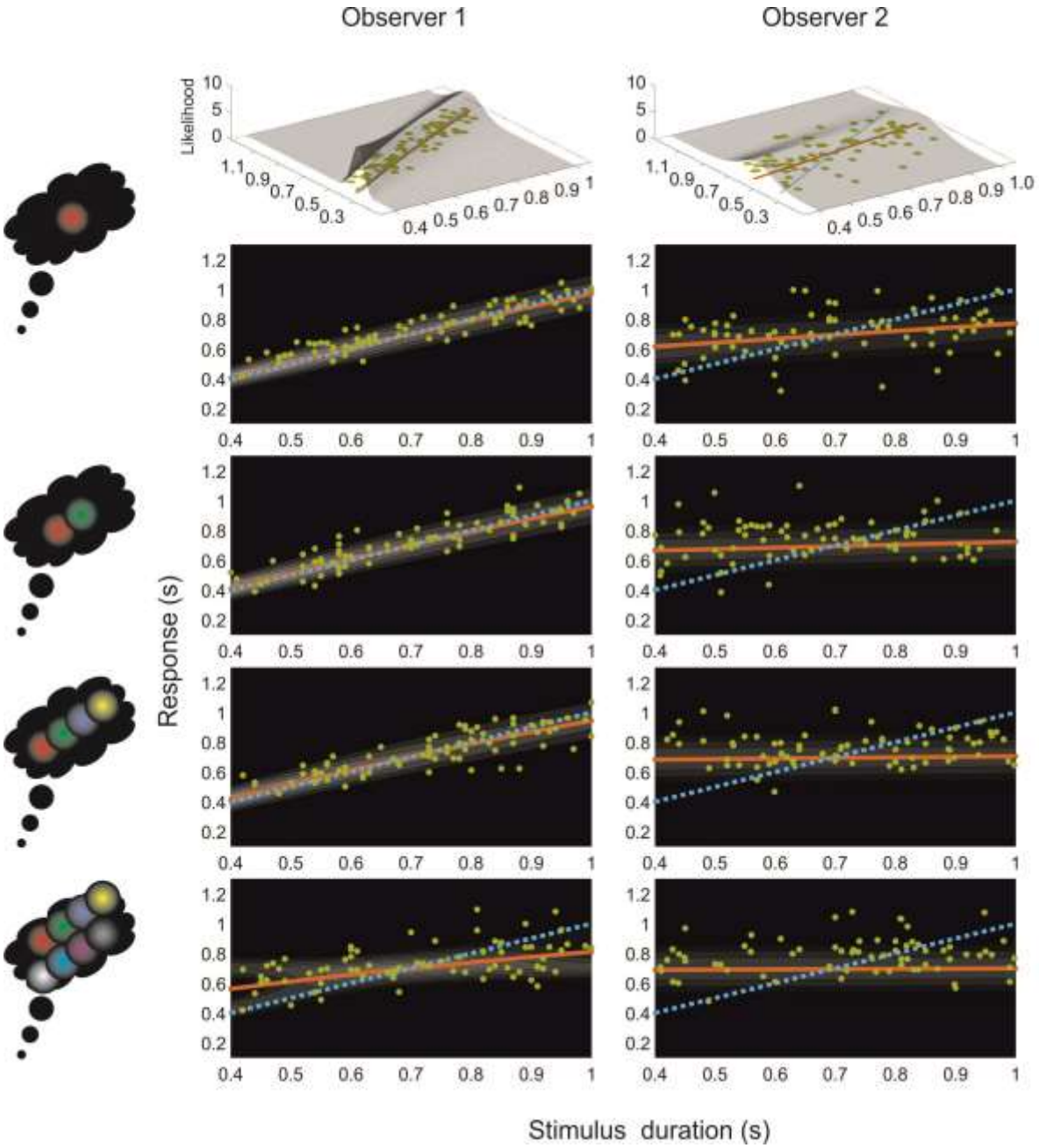

Figure A1. Model predictions and data from visual conditions for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both of these observers. This model also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 3, for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Best-fitting visual parameters for observer 1: Sensory Weber ratio 0.058; motor Weber ratio 0.096; noise exponent 0.305; slot capacity 3.963. Best-fitting auditory parameters for observer 2: Sensory Weber ratio 0.414; motor Weber ratio 0.192; noise exponent 0.332; slot capacity 0.953.

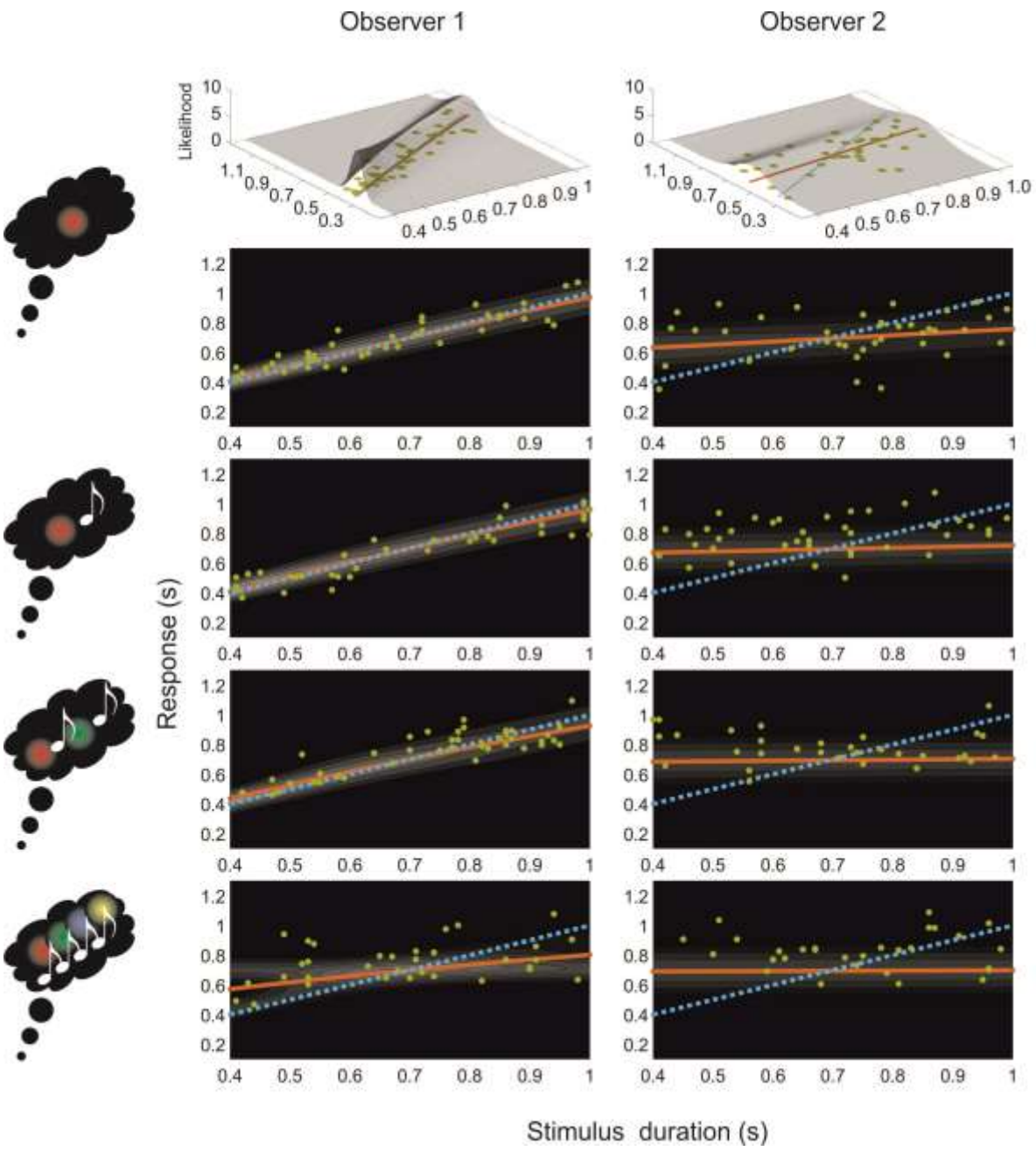

Figure A2. Model predictions and data from visual judgements made in intermixed conditions, for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both observers, and which also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 3 for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Model predictions vary very slightly from those shown in Figure A1 because load is affected by the intermixed load denominator parameter (observer 1: 0.934; observer 2: 0.862).

Appendix B

The validity of our model-based analysis was investigated via four parameter recovery simulations. In each, we simulated two complete sets of data for 30 experimental participants using one of our models. In the first set of data, each participant was simulated with d = 1 (i.e. a single unitary working memory store). In the second set, d was set to 2 (i.e. separate modality-specific stores). Other parameters were uniform random across plausible ranges (~0-0.5 for W_s and W_m, with a correlation between auditory and visual parameters, and W_{sAud} an average factor 1.5 greater than W_{sVis}; 0-8 for each n; integer 1-8 for each c).

Guessing from prior combined with late Bayes integration model simulated with 288 trials per participant

We first simulated the *Guessing from prior combined with late Bayes integration* model, with 288 trials per participant. This model is used for illustration in the main text as it achieved the lowest group-level AIC and also performed well for our two observers. The results of the parameter recovery, with identical trial numbers to those used in our main experimental group, are shown in Figure A3. The upper panel illustrates recovery of the all-important d parameter, which was set at either 1 or 2 for two simulated groups. It is clear that this parameter can be recovered successfully on average, but with some individual failures of recovery tending to generate positive skew across the group. Further investigation revealed that errors in recovery for d were uncorrelated with errors in recovery for most other parameters, but did correlate positively with errors in recovery for W_s parameters (data not shown), suggesting some possible trade-off. However, this issue does not appear to undermine the use of the recovered d estimates for inference at the group level.

Lower panels of Figure A3 show, within each simulated group, correlations between actual and recovered parameters for the remaining eight model parameters. Weber ratios are recovered well.

Parameters affecting how performance decreases with increasing load are recovered less well, and show some degree of clustering at extremes of the search range. The correlation matrix across these parameters for errors in recovery (data not shown) did not reveal any correlations consistently across both *d* = 1 and *d* = 2 groups, except where this was imposed by our generation and fitting procedures (i.e. between sensory and motor Weber ratios). This suggests that parameters do not tend to trade off (at least when considered pairwise).

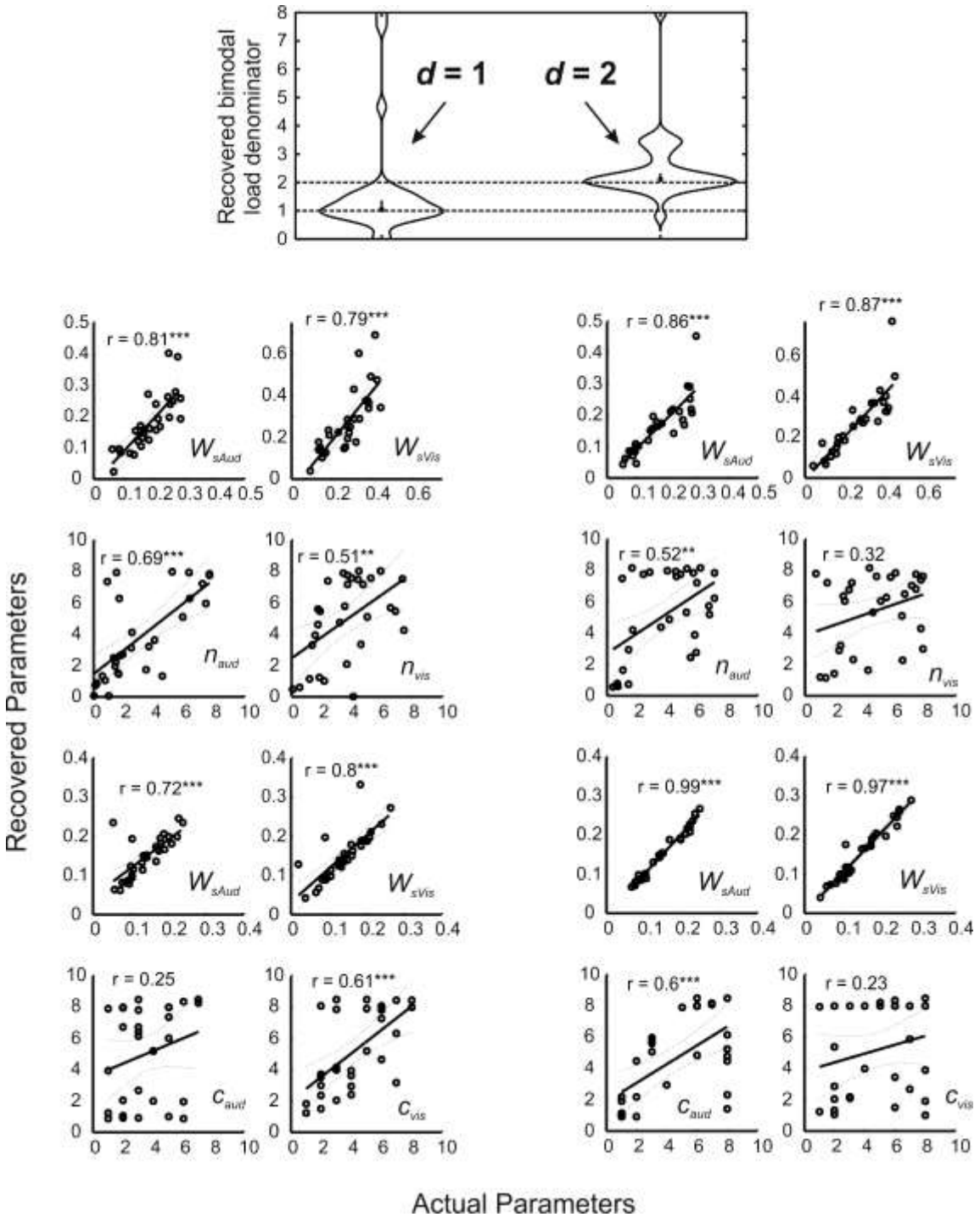

Figure A3. Results of parameter recovery simulation for guessing from prior combined with late Bayes integration model simulated with 288 trials per participant. Simulations were performed for two groups, one with d = 1 and one with d = 2 (top). The widths of black outline shapes represent probability densities. Medians and 95% bootstrap confidence intervals around medians are also shown. Other parameters were free to vary in each group (bottom). Black lines show best fitting linear prediction of recovered parameters from actual parameters. Dotted lines show 95% confidence interval around this fit.

A further three parameter recovery simulations varied the number of simulated trials (to match either the main experiment, or the observers, from Experiment 1) and the model (testing another plausible model, but a simpler variant with only one kind of parameter affecting how performance varied with memory load). Results from all four simulations are summarised in Tables A1 and A2, which show the median recovered value of *d*, and correlations between actual and recovered parameters, for simulated *d* values of 1 and 2 respectively. As might be expected, parameter recovery (as indexed by correlations) appears slightly better with a larger number of trials, and when only one kind of parameter controls how performance decrements with increasing load.

Table A1. Summary of model recovery simulations for two models each simulated with two trial counts, with simulated *d* = 1. IQR = interquartile range.

Model name	**Actual *d***	**Simulated Trials**	**Recovered *d* (IQR)**	*Correlations (actual vs. recovered)*							
				W_{sAud}	W_{sVis}	W_{mAud}	W_{mVis}	n_{aud}	n_{vis}	c_{aud}	c_{vis}
Guess from prior, combined with late Bayes	1	288	1.05 (0.64)	0.81	0.79	0.72	0.8	0.69	0.51	0.25	0.61
Guess from prior, combined with late Bayes	1	864	1.06 (0.31)	0.93	0.89	0.97	0.78	0.8	0.66	0.64	0.54
Memory noise with late Bayes integration	1	288	1.09 (0.63)	0.84	0.9	0.97	0.95	-	-	0.57	0.52

Memory noise with late Bayes integration	1	864	1.00 (0.14)	0.92	0.91	0.83	0.99	-	-	0.79	0.64

Table A2. Summary of model recovery simulations for two models each simulated with two trial counts, with simulated *d* = 2. IQR = interquartile range.

Model name	**Actual *d***	**Simulated Trials**	**Recovered *d* (IQR)**	*Correlations (actual vs. recovered)*							
				W_{sAud}	W_{sVis}	W_{mAud}	W_{mVis}	n_{aud}	n_{vis}	c_{aud}	c_{vis}
Guess from prior, combined with late Bayes	2	288	2.11 (0.80)	0.86	0.87	0.99	0.97	0.52	0.32	0.6	0.23
Guess from prior, combined with late Bayes	2	864	2.25 (0.8)	0.9	0.94	0.9	0.77	0.79	0.52	0.35	0.63
Memory noise with late Bayes integration	2	288	2.00 (0.53)	0.81	0.8	0.95	0.99	-	-	0.45	0.6
Memory noise with late Bayes integration	2	864	2.01 (0.36)	0.93	0.93	0.98	0.96	-	-	0.65	0.83

Tables

Table 1. Parameter summary for eight models of duration reproduction under memory load. See main text for description of how these parameters come to affect performance across the 3x4 experimental conditions.

Model name	**Number of parameters**	**Parameters handling single-item performance: Weber ratios.**	**Parameters handling load-induced performance decrements: Capacity and/or noise exponents.**	**Parameter handling changes to performance in intermixed conditions: Load denominator**
Slots plus guess from prior	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis}	d
Slots plus guess at mean of prior	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis}	d
Memory noise with late Bayes integration	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	n_{aud} , n_{vis}	d
Memory noise with early Bayes integration	7	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	n_{aud} , n_{vis}	d
Guess from prior, combined with late Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d
Guess from prior, combined with early Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d
Guess at mean of prior, combined with late Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d
Guess at mean of prior, combined with early Bayes	9	W_{sAud} , W_{sVis} , W_{mAud} , W_{mVis}	c_{aud} , c_{vis} , n_{aud} , n_{vis}	d

Table A1. Summary of model recovery simulations for two models each simulated with two trial counts, with simulated d = 1. IQR = interquartile range.

Model name	**Actual *d***	**Simulated Trials**	**Recovered *d* (IQR)**	*Correlations (actual vs. recovered)*							
				W_{sAud}	W_{sVis}	W_{mAud}	W_{mVis}	n_{aud}	n_{vis}	c_{aud}	c_{vis}
Guess from prior, combined with late Bayes	1	288	1.05 (0.64)	0.81	0.79	0.72	0.8	0.69	0.51	0.25	0.61
Guess from prior, combined with late Bayes	1	864	1.06 (0.31)	0.93	0.89	0.97	0.78	0.8	0.66	0.64	0.54
Memory noise with late Bayes integration	1	288	1.09 (0.63)	0.84	0.9	0.97	0.95	-	-	0.57	0.52
Memory noise with late Bayes integration	1	864	1.00 (0.14)	0.92	0.91	0.83	0.99	-	-	0.79	0.64

Table A2. Summary of model recovery simulations for two models each simulated with two trial counts, with simulated d = 2. IQR = interquartile range.

Model name	**Actual *d***	**Simulated Trials**	**Recovered *d* (IQR)**	*Correlations (actual vs. recovered)*							
				W_{sAud}	W_{sVis}	W_{mAud}	W_{mVis}	n_{aud}	n_{vis}	c_{aud}	c_{vis}
Guess from prior, combined with late Bayes	2	288	2.11 (0.80)	0.86	0.87	0.99	0.97	0.52	0.32	0.6	0.23
Guess from prior, combined with late Bayes	2	864	2.25 (0.8)	0.9	0.94	0.9	0.77	0.79	0.52	0.35	0.63
Memory noise with late Bayes integration	2	288	2.00 (0.53)	0.81	0.8	0.95	0.99	-	-	0.45	0.6
Memory noise with late Bayes integration	2	864	2.01 (0.36)	0.93	0.93	0.98	0.96	-	-	0.65	0.83

Figure Legends

Figure 1

Schematic of the experimental task and design. A trial is shown in which memory is loaded with two items, but load was randomly varied within each block (between 1, 2, 4 or 8 items). The white musical notes represent auditory stimuli and were not shown on screen.

Figure 2

Schematic of a Bayesian observer model for duration reproduction (Jazayeri & Shadlen, 2010). Responses are traced through the model pipeline for two example stimuli, one of short duration (400 ms; blue) and one of long duration (1000 ms, red). Across multiple trials stimuli with these objective durations are corrupted by scalar sensory noise, to yield a range of sensory point estimates (with Gaussian distributions, as shown in the bottom-left plot, labelled 1). From each such point estimate, the Bayesian brain would use knowledge about the noise-generating process to infer likelihood functions (two examples are shown in the top-left plot, labelled 2). Likelihood functions are combined with priors: The top-middle plot (3) shows both the uniform distribution of stimulus durations used in the experiment, and the (assumed) Gaussian prior (which is derived from this uniform distribution). A point estimate is then recovered, based on the mean of the resulting posterior distribution (top-right plot; 4). Finally, these perceived durations are targeted, but further corrupted by scalar motor noise during reproduction, to yield the distribution of responses shown in the bottom-right plot (5).

Figure 3

Illustrative predictions for four models of short-term memory for duration. Parameters used in simulations were: Sensory weber ratio = 0.1, motor weber ratio = 0.1, and either memory slot capacity = 2 or noise exponent = 0.5. Thin or dashed grey lines show objectively correct

performance, i.e. reproduced duration = stimulus duration. Light grey 3D overlay or heat maps show predicted probability densities for responses made to stimuli with different durations. Thick black lines show the means of these predictions. A. With only a single item in memory, all four models make an identical prediction, reflecting the Bayesian-observer model schematised in Figure 2. For a high-precision observer, the biasing effect of the prior is subtle (but slightly more evident at higher durations that accrue greater sensory noise) and mean performance is close to being objectively correct. B. When memory load is increased to eight items, performance is worse, and predictions vary for different models. From top to bottom: The slots plus guess from prior model shows highest probability density near the objectively correct value, but also a plateau region extending across the prior, which reflects guessing when memory capacity is exceeded (this dramatically affects the mean response); second plot down – the slots plus guess at mean of prior model shows greatest probability density near the mean of the prior (as with eight items in memory, and a two-item capacity, guessing is the most likely outcome) but also a marked probability of responding quite accurately (when the item has been remembered). This results in intermediate values for the mean response; third plot down – for the memory noise with late Bayes integration model, the increased noise in all memorised items yields a greater influence of the prior during the Bayesian integration process, and thus a shallower slope relating stimulus duration to mean reproduced duration; finally, fourth plot down – in the memory noise with early Bayes integration model, memory noise is added only after the Bayesian integration process, so the influence of the prior remains limited and mean performance shows little bias. There is, however, a variable error that increases dramatically relative to the single item case (note the shallower profile relative to the single case, depicted on the left).

Figure 4

Group average data. Error bars denote standard error of the mean. **A.** Average root mean squared errors (RMSE) of interval reproduction across 30 participants for auditory and visual stimuli, presented in either unimodal or intermixed blocks, and for memory loads of 1 to 8 items. **B.** Average

RMSE of interval reproduction stratified by position of target within stimulus sequence and memory load (but collapsed across stimulus modality). **C.** Mean reproduced intervals in each condition. Prior to averaging across participants, responses on individual trials were smoothed into a moving average using a Gaussian kernel (cf. Kohl et al., 2019).

Figure 5

Mean-centred AIC (left) and BIC (right) values for best-fitting variants of eight models of short-term memory for duration. Models 1-4 implement either slot limits or memory noise to explain performance decrements with increasing load, whereas models 5-8 cross combine these two kinds of modelling approach. The mean value across models for each participant has been subtracted from their scores to increase y-axis resolution. Upper plots show data for the two observers who completed additional blocks for illustrative purposes. Lower plots show data for a further 30 participants (grey lines), along with their group average scores (black lines) and 95% confidence intervals (error bars). Winning models at the group level are underlined. Models with significantly higher AIC/BIC values than these winning models (Tukey's honestly significant $p < 0.05$) are denoted by grey text and asterisks (*).

Figure 6

Model predictions and data from auditory conditions for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both of these observers. This model also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 2, for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model

predictions. Conditions with increasing memory loads are presented further down the figure. Best-fitting auditory parameters for observer 1: Sensory Weber ratio 0.053; motor Weber ratio 0.070; noise exponent 0.640; slot capacity 4.827. Best-fitting auditory parameters for observer 2: Sensory Weber ratio 0.181; motor Weber ratio 0.120; noise exponent 0.174; slot capacity 1.084.

Figure 7

Model predictions and data from auditory judgements made in intermixed conditions, for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both observers, and which also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 2 for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Model predictions vary very slightly from those shown in Figure 6 because load is affected by the intermixed load denominator parameter (observer 1: 0.934; observer 2: 0.862).

Figure 8

Kernel density plots summarising group parameter values from best-fitting models. The widths of black outline shapes represent probability densities. Red markers indicate medians. Red vertical bars indicate 95% bootstrap confidence intervals around medians. For intermixed load denominator parameter plots, dashed horizontal lines at values of 1 and 2 indicate predictions for single-store and dual-store accounts, respectively. WR = weber ratio. **A.** All nine parameters are illustrated for the well-supported *guess at mean of prior, late Bayes integration* model. The intermixed load

denominator parameter does not differ significantly from 1.0, but does differ significantly from 2.0, supporting a single-store account. **B.** The intermixed load denominator parameter for all five plausible models (see Figure 4). Confidence intervals rarely overlap 2, but almost always overlap 1, again supporting a single-store account.

Figure A1

Figure A1. Model predictions and data from visual conditions for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both of these observers. This model also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 3, for the load 1 condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Best-fitting visual parameters for observer 1: Sensory Weber ratio 0.058; motor Weber ratio 0.096; noise exponent 0.305; slot capacity 3.963. Best-fitting auditory parameters for observer 2: Sensory Weber ratio 0.414; motor Weber ratio 0.192; noise exponent 0.332; slot capacity 0.953.

Figure A2

Figure A2. Model predictions and data from visual judgements made in intermixed conditions, for two observers (left/right) who completed additional blocks for illustrative purposes. Model predictions come from the guess at mean of prior, late Bayes integration model, which was among the better models for both observers, and which also performed well at a group level (see Figure 5). In the uppermost plots, model predictions are presented in the format of Figure 3 for the load 1

condition. The same data and predictions are presented one row down, but using a 2D shaded-contour format to aid visualisation of data points, with lighter background shading denoting higher predicted probability densities. In both formats, dashed blue lines denote objectively correct responses, and solid red lines indicate mean model predictions. Conditions with increasing memory loads are presented further down the figure. Model predictions vary very slightly from those shown in Figure A1 because load is affected by the intermixed load denominator parameter (observer 1: 0.934; observer 2: 0.862).

Figure A3

Results of parameter recovery simulation for *guessing from prior combined with late Bayes integration* model simulated with 288 trials per participant. Simulations were performed for two groups, one with $d = 1$ and one with $d = 2$ (top). The widths of black outline shapes represent probability densities. Medians and 95% bootstrap confidence intervals around medians also also shown. Other parameters were free to vary in each group (bottom). Black lines show best fitting linear prediction of recovered parameters from actual parameters. Dotted lines show 95% confidence interval around this fit.