Revision of XGE-2021-3289 as invited by the action editor, Adam Waytz.

Supplemental Material for:

Background Music Changes the Policy of Human Decision-Making: Evidence from Experimental and Drift-Diffusion Model-Based Approaches on Different Decision Tasks

Agustín Perez Santangelo[1,2], Casimir J.H. Ludwig[3], Joaquín Navajas[2], Mariano Sigman[2], and María Juliana Leone[2,4]

[1]Instituto de Investigación en Ciencias de la Computación, Universidad de Buenos Aires, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)

[2]Laboratorio de Neurociencia, CONICET, Universidad Torcuato Di Tella

[3]School of Psychological Science, University of Bristol

[4]Laboratorio de Cronobiología, CONICET, Departamento de Ciencia y Tecnología, Universidad Nacional de Quilmes

Supplemental Text

Timing Hypothesis Control Analyses

We performed two control analyses to assess the *timing* hypothesis in more detail. Specifically, according to this hypothesis, the observed reduction in RTs for both slow and fast music should correlate with a *tempo*-dependent arousal increase.

First, we verified that participants were able to synchronize to the beat of the slow and fast music in order to measure *tempo* perception (John Iversen, 2008). To this end, we evaluated participants ability to tap to the beat of the music that was played during the decision tasks -at the end of the study- and quantified their performance using circular statistics. We found that participants were generally able to synchronize to the tempo of the music (Figure. S2), showing that the tempo of the music was effectively perceived. To ascertain that the perceived faster *tempo* induced higher subjective arousal as expected (Husain et al., 2002), we collected subjective arousal ratings (in arbitrary units (a.u.), ranging from 0 to 1 with 0.001 resolution) for each music track at the end of the study. We fit the data with a zero-one-inflated-beta generalized linear mixed model (Arousal-rating-ZOIB, see Supplemental Methods for model specification) and found that fast music induced higher arousal levels than slow music ($\overline{\Delta rating}_{(fast\text{-}slow)}$ = 0.30 a.u., CI_{95} = [0.21, 0.38]) (Figure. S3a).

Second, we assessed whether music *tempo* affected arousal states *during* decision-making (measured at the end of each trial-block). We fit arousal data with a ZOIB-GLMM (Arousal-ZOIB) which revealed that, across-tasks, fast music induced higher arousal than silence ($\overline{\Delta arousal}_{(fast\text{-}silence)}$ = 0.11 a.u., CI_{95} = [0.06,0.15]) and slow music ($\overline{\Delta arousal}_{(fast\text{-}slow)}$ = 0.06a.u., CI_{95} = [0.01,0.11]), which also induced higher arousal than silence ($\overline{\Delta arousal}_{(slow\text{-}silence)}$ = 0.05a.u., CI_{95} = [0.01, 0.08]) (Figure. S3b). Thus, participants' subjective arousal was

increasingly higher with faster music *tempo*, relative to silence, throughout all tasks. But did this *tempo*-dependent arousal induction translate to a monotonical reduction of RT, as predicted by the timing hypothesis? To answer this, we modeled mean RT for correct responses with a shifted log-Normal GLMM (Arousal-RT-GLMM) that included a continuous predictor for subjective arousal. By the timing hypothesis we expected a negative coefficient for this predictor, i.e., faster RT for higher arousal. The model revealed that overall ($\overline{\beta}_{arousal}$ = -0.04, CI_{95} = [-0.13, 0.05]) and, consistently, by-task arousal levels were not significantly related to RT (Table S2).

Beat-Synchronization

We analyzed whether participants may have used the music beat as a cue to respond, relying on a motor-synchronization strategy to decide. By this idea, participants would press the key to respond when they expected the beat of the background music. Indeed, as described above, participants were generally able to synchronize to the beat of the music played during the decision tasks (Figure S2). We analyzed whether they relied on this strategy to decide, using circular statistics to assess to what extent the keypresses were phase-locked to the background-music-beat times (see Supplemental Methods). We found that only 10% of the keypress-time data was compatible with a synchronization strategy (even allowing for a non-zero offset, i.e., responses not perfectly aligned to the beat), but not consistently either within participants, task, or music condition (e.g., a participant that synchronized on RDK trials with slow music, did not with fast music in the same task) (Figure S4). This suggests that synchronization did not play a relevant role in the effects of music on decisions.

Liking

After listening to a 12-second excerpt of each musical track (the same we used as our experimental manipulation), participants rated their subjective liking with a clickable visual

analogue scale (ranging from 0 -*nothing*- to 1 -*a lot*- with a resolution of 0.001). We visually inspected these data and found that it did not meet any informative distributional assumptions. Thus, we performed a non-parametric paired-sample Wilcoxon signed rank test for paired samples over the difference in liking between music conditions for each participant (using Bayesian methods with *bayesWilcoxTest (Reinhardt, 2020)*). This test revealed that both conditions produced similar liking ratings as the difference ($\Delta like_{(fast\text{-}slow)}$ = -0.11 a.u, CI_{95} = [-0.54, 0.33]) was not significantly different from zero. Further, we performed a similar test to assess whether ratings were lower than the middle point of the visual-analog scale (i.e., 0.5). We found that liking ratings were lower for both slow (median = 0.40, V = 159.5, p-value = 0.026) and fast music (median = 0.35, V = 110, p-value = 0.0035). These results held when grouping participants by level of musicianship. For musicians, $\Delta like_{(fast\text{-}slow)}$ = -0.02 a.u. (CI_{95} = [-0.77, 0.70]) and both liking ratings were lower than the scale middle point (slow music: median = 0.34, V = 23, p-value = 0.062; fast music: median = 0.32, V = 18, p-value = 0.03). For non-musicians, $\Delta like_{(fast\text{-}slow)}$ = -0.16 a.u. (CI_{95} = [-0.70, 0.37]) and both liking ratings were lower than the scale middle point (slow music: median = 0.42, V = 63.5, p-value = 0.11; fast music: median = 0.37, V = 42, p-value = 0.03).

To provide reassurance to these findings, we employed a parametric approach (fitting data with a ZOIB-GLMM, similar to arousal-ratings analysis) and found consistent results.

Additional Music Stimuli Testing

First, to address the possibility that the slow tempo (40 bpm) manipulation could have distorted relevant features of the original music piece, we ran a reduced version of our experiment (with Marble task only) in which we tested the effects of a new slow version (at 70 bpm) on decision-making. We analyzed data the same way we did for the main experiment. We

observed (N = 20) that the new slow music had similar effects on decisions as the 40 bpm music had, i.e., it induced faster ($\overline{\Delta RT}_{(slow\text{-}silence)}$ = -124ms, CI_{95} = [-169, -87] ms) and less accurate ($\overline{OR}_{(slow/silence)}$ = 0.813, CI_{95} = [0.597, 1.110]) decisions, which mapped onto a lower decision threshold ($\overline{\Delta a}_{(slow\text{-}silence)}$ = -0.161, CI_{95} = [-0.225, -0.099]) in DDM (Figure S8).

These results suggest that the new slow version (which should be normatively less disruptive of temporal coherence) had similar effects as those we reported using the original slow music, providing further support for both the appropriateness of our stimulus selection and the discussed tempo-independency of the observed effects on decision-making.

Second, to understand -at least on an exploratory basis- to what extent our results and interpretations hold with other music, we ran an independent reduced version of our original experiment (with the Marble task only), but this time using new stimuli. We used for this new experiment a blues and a techno excerpt (both at 190 bpm). Our original results suggested that mood (instead of arousal) might be the main driver of the effects of music we observed. Thus, we expected (although there are certainly many factors that are different between our original stimuli and these new pieces, aside from pleasantness) that more pleasant music (like the blues excerpt, relative to techno) could induce slower and more accurate decisions. The results from this new experiment showed that blues music had minimal impact on decisions (slightly slower yet inaccurate decisions, mapping to a marginally more cautious decision threshold). In contrast, the techno excerpt, produced similar effects as those described in the main text: faster ($\overline{\Delta RT}_{(slow\text{-}silence)}$ = -41ms, CI_{95} = [-79, -2] ms) and less accurate ($\overline{OR}_{(slow/silence)}$ = 0.808, CI_{95} = [0.552, 1.170]) decisions, which mapped onto a lower decision threshold ($\overline{\Delta a}_{(slow\text{-}silence)}$ = -0.071, CI_{95} = [-0.143, -0.001]) in DDM (Figure S9).

Thus, it seems that there could be a shared signature between the original stimulus and this new techno excerpt that is not shared with the blues track. This signature might be an "unsettling" character, given that liking, pleasantness and tension ratings for these pieces were lower than the ratings for the *blues* excerpt. At the same time, the shy results with *blues* might be -in part- attributable to effect saturation (since accuracy rates are already high for silence, there is little room for the putative expected improvement in accuracy) and low number of trials and subjects (since we had to adapt our procedure to a non-in-lab setting). Interestingly, both new tracks (blues and techno) had the same fast tempo. This suggests that, consistent with the tempo-independency we observed with our original stimuli, *tempo* might have a secondary role in how music affects decisions (since music at the same tempo produced different effects on decisions), and rather mood might be the main driver.

Supplemental Methods

Participants Sample Size

We ran a small pilot study (n=8) with the same design as the main experiment (except that we tested only one decision-making task: Marble) in order to estimate an effect (β_{40bpm} = -31ms) that would allow us to calculate the appropriate sample size for the main experiment. Pilot data power analysis (based on 1000 iterated simulations (Brysbaert & Stevens, 2018; Green & MacLeod, 2016) per sample size analyzed, namely 8, 16, 24, 32 and 40) showed that -to detect a significant effect of music with a Wald z-test over the coefficient for slow music over RT- we would need, at least, n = 24 to achieve a power of at least 80%. Since our study required a 3-session commitment, we intentionally overshot the number of participants to 32 to safeguard sample size from any possible participant dropping out. With n=32 estimated power was 92.4% (95% confidence interval = [90.58, 93.97]), at significance threshold $\alpha = 0.045$.

Decision-Making Tasks

Tasks were programmed in MATLAB (MathWorks) using Psychtoolbox extensions (Brainard, 1997). Task structure is detailed in the main text methods. For all tasks, trials were interleaved by an inter-trial interval (ITI) drawn from a uniform distribution ITI ~ uniform (0.7s, 1s). During the ITI a fixation cross was displayed (.65º visual angle) to draw attention to the center of the screen. Next, stimuli were displayed on screen until a response was made or until a time deadline (3 seconds, learnt by the participant during practice trials) was reached. Overall, only 583 trials were lost due to time-outs (~0.58% of the total database).

Next, we provide details for each task:

Random dot kinetogram (RDK)

This perceptual-motion task was adapted from refs. (Shadlen & Kiani, 2013; Zylberberg et al., 2012). Participants had to decide whether a moving cloud of dots (each dot had size=0.1º and moved at speed =50º/s) presented in a centered circular aperture (8º diameter) appeared to move to the left or to the right. Difficulty was given by the proportion of dots moving in the same direction (i.e., % coherence) and we set it at 51.2, 12.8 and 3.2% (for each response location), representing easy, medium and hard trials, respectively. Dot density was set at 16.7 dots/$(º)^2$/s. Sequential position of dots was determined by a limited-lifetime algorithm(Palmer et al., 2005) to avoid single-dot tracking: A first group of dots is shown in the first frame, a second group are shown in the second frame, a third group is shown in the third frame. Then in the next (4th) frame, some percentage of the dots from the first frame are replotted in motion according to the speed/direction and % coherence value (as opposed to randomly replaced). Similarly, the same is done for the second group, etc.

Marble

This numerosity and contrast-based perceptual task was adapted from ref.(Dutilh & Rieskamp, 2016) . Participants had to decide whether a screen-centered 7.72º-side square board (displayed over a gamma-corrected gray background (to avoid any confounding effects of luminance non-linearity)) containing 100 black and white marbles in total (0.45º diameter) had a greater number of black or white marbles. Difficulty was given by the percentage of white marbles and we set it at 38 (62), 45 (55) and 48 (52) for black (white) trials, representing easy, medium, and hard trials, respectively.

City

This inference-from-memory general knowledge task was adapted from ref.(Pleskac & Busemeyer, 2010) and comprised two consecutive stages. First, we screened the knowledge of 82 world cities by sequentially showing the participants the names of these cities and asking them to report whether they knew each of them with a keypress. Then, for each participant, we constructed all possible city pairs from the reportedly known cities and computed the absolute difference between the official population of both. Then, to set difficulty levels, we divided the whole set of pairs into tertiles following that difference value and flipped city-names according to the response location. On the second stage, the names of the cities in each pair were displayed in yellow (Arial 12) to the left and right of the screen-center (6.2º eccentricity) and participants had to choose the one –left or right- they considered had the highest population. No city name was repeated on consecutive trials.

Lexical

This lexical categorization task was adapted from refs. (Gonzalez-Nosti et al., 2014; Ratcliff et al., 2004). Participants had to decide whether a string of 5 to 8 letters displayed in white (Arial 35) on the center of the screen was a word in Spanish or not. Following ref.(Ratcliff et al., 2004), we constructed difficulty levels over a *wordness* measure which integrates several lexical features that map onto evidence-accumulation rates in the Drift-Diffusion Model of decision-making (DDM). First, we downloaded (from the EsPal Corpus(Duchon et al., 2013)) and annotated 2605 Spanish 5-to-8-letter nouns (with all lexical-features metadata), with RT and age of acquisition (AoA) data from ref(Gonzalez-Nosti et al., 2014). Since faster accumulation rates yield faster and more accurate lexical categorization, we explored how lexical variables correlated to these decision outcomes. We found that -qualitatively- higher log frequency of occurrence, lower AoA and lower orthographic Levenshtein distance 20 (OLD20, i.e., the minimum number of edits required to obtain 20 words from a given word) were associated with faster RT. Thus, we used RT as a summary measure of *wordness.* We divided all words into 3 groups according to RT terciles, which represented easy, medium and hard difficulties. We randomly divided each group into two subsets (we checked all lexical features (including distribution of word length) were comparable between subsets). One subset was used for word stimuli, the other was used as a base-word pool to construct non-words. Non-word candidates were obtained using the Wuggy algorithm (Keuleers & Brysbaert, 2010) (which preserves language-specific phonotactic features) over each base-word string. We selected the non-word candidate that minimized the difference between its OLD20 and the base-word's OLD20. Then, we assigned the difficulty level for these non-words by inverting the difficulty of the base-words it was derived from (e.g*.,* easy -low RT- base-words was categorized as a hard non-word). This way, we obtained a database comprising easy, medium, and hard words and

non-words, from which we sampled 120 of each category (720 in total) to use as our stimuli pool for each task run.

Snack.

This value-based task was adapted from ref.(Milosavljevic et al., 2010), and comprised two stages. First, participants reported their subjective ratings (i.e., how much they would like to eat) for each of 100 snacks using a 5-point Likert scale ranging from -2 (*nothing*) to 2 (*a lot*) starting point was random to account for anchoring effects. Before participants started to rate, they watched all snacks on screen to promote an effective use of the scale. Once we collected all snack ratings, we constructed snack pairs according to their rating difference. On the second stage, snack-images of a pair were displayed to the left and right of the screen center (4.9º eccentricity) and participants had to choose the item they preferred. A 6.5º-sided yellow squared box was displayed around the selected snack. Difficulty was given by the absolute difference between subjective values reported on the first stage for each item of a pair and we set it at 3, 2 and 1 for E, M and H trials, respectively for each response location. No given item was shown in two consecutive trials. Further, no accuracy feedback was given throughout the task. All images were scraped from a popular Argentinian supermarket online-catalogue and then embedded on a black background for display.

Auditory stimuli manipulation

We passed a piano VSTi (Pianoteq 6 stage) to a Musical Instrument Digital Interface (MIDI) file of the composition originally adapted for piano and added a percussive quarter-note beat track to emphasize *tempo* as a salient feature. Care was taken to prevent audio deformations and to assure even levels of sound pressure for both tracks (approx. 60dB) during

playback. Both tracks contain only filled intervals (i.e., there are no silences nor gaps) which is important in light of possible time-distortions when judging filled and unfilled intervals (e.g., clicks) (Wearden et al., 2007). Finally, high-quality (sample rate = 44.1 kHz, bitrate = 352kbps) mono Wave files (available at https://osf.io/fguq6/) were exported and later played (on both, left and right channels) when called by the task-scripts, via PsychPortAudio (a sound driver from the Psychtoolbox suite).

Music-related tasks

The tapping task was adapted from ref.(John Iversen, 2008). After practice with unrelated music, participants were instructed to follow the beat of the music that was going to be played, with consecutive keypresses (in the same fashion they would usually tap with their feet or hands). We collected 50 keypresses per music track.

Arousal and liking ratings were measured after the participant listened to a 12-second excerpt of each track, with a clickable visual analogue scale ranging from 0 (*nothing)* to 1 (a *lot),* with resolution of 0.001. Importantly, as for arousal reports during the decision tasks, we instructed participants to report *felt* arousal, and we used the same prompting question.

Music order was randomized between participants for both tasks. Tapping ability and ratings for liking and arousal were measured at the end of the experiment to avoid revealing the purpose of the study to participants prematurely.

Data Analysis

Decision outcomes (RT and accuracy)

Our statistical approach relied on Bayesian estimation of generalized linear mixed models (GLMM) with Hamiltonian Monte Carlo (HMC) sampling method. GLMMs extend the scope of general linear models to all the exponential family of probability distributions.

Moreover, random effects modeling (the “mixed” part of GLMM) allowed us to represent data-dependency through grouping variables which, at the same time, allowed for population-level inference and interpretation, and increased model parsimony (less parameters to estimate). Further, this strategy is robust to unbalanced data (as was ours). We created our GLMMs within the Stan computational framework (http://mc-stan.org/) accessed with *brms* package (Bürkner, 2017) in R (R Core Team, 2020). To improve convergence and guard against overfitting, we specified weakly informative priors.

We first excluded timed-out (~0.58% of all data) and impulsive responses (RT < 0.25sec (< 0.2% of all data)) from the database since these data do not represent decisions.

For the RT analysis, we used RT for correct trials because correct responses better represent decisions and because error trials were slower (median $RT_{correct}$ = 810 ms, median RT_{error} =1150 ms), only represented 16.4% of data and were not homogeneously distributed across tasks (RDK 3.1%, Marble 2.7%, City 6.4%, Lexical 1.5%, Snack 2.7%), which hindered model estimation.

RT conditional distributions (by task and by music condition) were right-skewed (as is typical for these decisions (Ratcliff & McKoon, 2008)), and observed conditional (by participant, task and music condition) means were proportional to its dispersion (standard deviations, SD). Thus, we fit our data with a shifted log-normal (Wagenmakers & Brown, 2007) GLMM. Since our approach was hypothesis-driven we defined our minimal model as the one including the hypothesis-related fixed factors (music and task) and their interaction, and by-participant random effects over these fixed factors, for the distribution location parameter (μ).

We then specified, fit, and compared growingly complex models that included (or not) control-related fixed factors (and their interactions) and that varied in their random-effects structure. Regarding the fixed factors, we included difficulty, response location, their interaction,

their interaction with task, their interaction with music condition and their triple interaction with task *and* music condition. Further, given that each task was completed in approx. 25min, chronological effects (fatigue, learning, hurry-to-leave) were likely to play a role in observed performance (Parasuraman & Mouloua, 1987). However, since music-condition order within a task was not fixed, music and chronology effects were not confounded. Thus, we included a mean-centered scaled predictor for trial number (and its interaction with task) to control for this effect. We also varied the random-effects structure, by including (or not) order-related grouping variables (task order and music order within a task) and by-participant variability over fixed-effects (random slopes). For each of these growingly complex models, we drew 6000 samples (four chains of 2500 samples from which 1000 were for warmup) with an adaptive delta of 0.95 to minimize non-divergent transitions after warmup.

To diagnose each model, we visually inspected MCMC chains with trace plots for each parameter, we evaluated their autocorrelation, and assessed chain-mixing by determining if $\hat{R}$ values were smaller than 1.01 using the *launch_shinystan* function (from *rstanarm* (Goodrich et al., 2018)).

We then performed selection by computing and comparing the out-of-sample predictive accuracy of the models with an approximation to leave-one-out (loo) cross-validation using Pareto smoothed importance sampling (Goodrich et al., 2018). The model with the highest predictive accuracy was selected as the winning model (see its specification in main text Methods). For this model, we then drew 12000 samples (four chains of 5000 samples from which 2000 were for warmup) with an adaptive delta of 0.95.

We validated the winning model with posterior predictive checks (PPC) by simulating 2000 new datasets from the estimated parameters posterior distributions, and computing mean and SD of the conditional distributions (by task and music) of each set. We then obtained the

mean and credible interval (containing 95% of the probability density, CI_{95}) of the distribution of mean and SD values and compared it to the observed-data summaries. This approach not only allowed us to assign credibility to the model, but also to have a measure of uncertainty derived from the distribution of simulated summary statistics. We also computed a Bayesian approximation to R^2 (18) for a point estimate of the model (the mode) to have a more typical sense of in-sample goodness-of-fit.

For the accuracy analysis, we applied the same method as for RT, except that here the dependent variable is binary (i.e., it only takes two values 1: correct response, 0: error) and was accordingly modeled as a random variable with a Bernoulli probability distribution with a *logit* link to the linear predictor.

For validation, additionally to PPC and instead of R^2, we computed the area under the Receiver Operating Characteristic curve (AUC-ROC), which reflects the classifying ability of the model predictions (as a rule-of-thumb, AUC-ROC = 0.8 denotes a good predictive model).

Finally, for both models, we tested our hypotheses directly over the posterior (i.e., updated) probability distribution of an effect (i.e., difference/odds ratio between marginal or conditional means), by computing the CI_{95}, (which is the benchmark for hypothesis testing within the Bayesian framework) of the effect distribution and assessing whether this interval contained a null-effect value (0 for differences, 1 for odds ratio). If the CI_{95} did not include this value, we interpreted this as credible support for the hypothesis that the corresponding effect is different from that null value, which we refer in the main text as "*significant*", to match the concept from the more traditional frequentist framework. Marginal and conditional posterior distributions were computed either with custom R code or with the *tidybayes* package (Kay, 2020). To aid interpretation, the summary values of the reported effects (mean and CI_{95} for both RT and

accuracy) were calculated after transforming the marginal/conditional posterior distributions to the appropriate scale: milliseconds (ms) for RT; odds of a correct response for accuracy.

Subjective arousal

We divided the analysis in three parts and used the same overall rationale (model diagnosis, selection, validation, and hypothesis testing) as for the decision outcomes analysis.

First, to determine the effect of music *tempo* on subjective arousal we fit arousal ratings with a zero-one-inflated-beta (ZOIB) (Ospina & Ferrari, 2012) GLMM using Bayesian methods. This mixture-of-distributions modelling approach is a better representation of the putative underlying generative process of arousal ratings (bounded between 0 and 1) and allowed us to dissect discrete events (zeros and ones) from continuous ratings and build linear models over each of its four parameters. After selection, we found that the model with music condition (slow, fast) as fixed factor and with random intercepts by participant on μ (mean of re-parameterized beta distribution) explained our data best.

Similarly, to determine the effect of music *tempo* on subjective arousal *during* decision-making we fit arousal ratings with a ZOIB GLMM with music condition, task, their interaction and a numerical predictor for block number (to account for chronological effects (van den Brink et al., 2016), allowed to vary by task) as fixed factors and full random-slopes by participant ID.

Third, to test the relation between arousal and RT, we computed the RT-mean of the last six correct trials by trial-block as our dependent variable (since arousal was measured at the end of each trial-block). We specified a shifted-logNormal Bayesian GLMM that included task and a continuous predictor for arousal (and their interaction) as fixed factors. We also included predictors to control for mean trial difficulty (in those six trials) and chronological effects (using

block number), and their interaction with task. We also specified Participant ID as grouping factor with random effects over all fixed factors (but not their interactions). Since arousal is a continuous predictor, we computed the marginal (across tasks) and conditional (by task) regression coefficients for arousal with their respective CI_{95} and tested our hypothesis by assessing whether these intervals contained 0 (Table S2).

Tapping

First, we removed the first 10 keypresses of each tapping session as a conservative measure against "landing" presses (i.e., adjusting to the beat progressively). Then, to understand whether participants were tapping on the beat, we built a beat vector (i.e., a vector with the actual-beat times. E.g., for slow music this vector is: [0, 1.5s, 3s, 4.5s, etc.]) and computed the difference of each keypress time to the closest beat in this vector. These differences were divided by the beat duration (1.5s for slow, 0.31s for fast) to obtain relative differences (by definition, maximal relative difference is half the beat duration) which were then transformed to radians. Next, we analyzed these data with circular statistics based on the circular version of the Normal distribution: the von Mises probability distribution. Specifically, we first tested unimodality using *CircMLE* (Fitak & Johnsen, 2017), which determines the best circular model (among models that include uniform, unimodal and bimodal circular distributions) and applies a Rayleigh test for unimodality that tells whether there is evidence for rejecting the hypothesis of a uniform distribution over the circle, but not whether there might be a phase-shift (i.e. a participant might have tapped at the actual beat, but with a constant delay of time). To assess whether participants were synchronizing their tapping phase to the beat, we tested whether the mode of these unimodal data was equal to zero with a V-test (Landler et al., 2018).

Last, since these analyses were uninformative of whether participants tapped over multiples of integer powers of 2 of the beat durations (e.g., for slow music, instead of tapping every 1.5s, every 0.75s or 3s), we computed the mean and standard deviation (SD) of the difference between consecutive keypresses. For clarity, we also computed the ratio of these means and SD to the beat duration. This way, the value of this beat ratio represented whether participants tapped at the actual *tempo* (ratio = 1), faster (ratio <1) or slower (ratio>1).

Beat-synchronization

To detect whether participants adopted a synchronization strategy to decide, i.e., they were pressing the key to respond when they expected the beat of the music that was being played on the background, we followed the same rationale as for the tapping analysis. However, here, we evaluated the alignment of the responses (decisions) to the background-music beat. Critically, we recorded the timestamp of each response and of the music onset (which was played continuously over 5 trial-blocks) within the decision-task MATLAB scripts. Thus, by constructing a beat-vector over that period, we were able to compute the difference of the responses' keypress times relative to the beat duration. After data transformation we applied the same circular modeling strategy as for tapping analysis and applied a Rayleigh test to detect distribution unimodality (Landler et al., 2018). Further, we computed the proportion of block-groups on which participants relative keypresses were indicative of synchronization (significant Rayleigh test) and we qualitatively assessed consistency of this strategy within participant, task, and music condition. Last, we considered whether participants that synchronized on more than one block-group of trials, had a consistent phase (i.e., whether they were responding in a specific moment of the beat-cycle). Homogeneity in this phase by participant would speak in favor of a synchronization strategy.

Decision Process (DDM)

To dissect the effects of background music on the decision process, RT and accuracy data were simultaneously fit to the DDM using the HDDM Python toolbox (Wiecki et al., 2013) which implements Bayesian estimation of parameters with literature-based priors following a generative-node hierarchical tree structure. This is a flexible and reliable version for DDM parameter estimation that is robust to unbalanced data (Ratcliff & Childers, 2015). Critically, we constructed linear (mixed) models over each DDM parameter using the *HDDMRegressor* function, to quantify and test the impact of our experimental conditions on each component of the decision process. We specified a minimal model which included main-effects terms for task, difficulty and chronological effects and was informed by theoretical constraints (Ratcliff & McKoon, 2008), namely: evidence threshold was allowed to vary between tasks, difficulty only affected accumulation rate (since participants are unaware of the type of trial they would get next, thus virtually dismissing any anticipatory strategy that would impact on the evidence threshold); starting point was fixed at 0.5 (unbiased priors); non-decision time was free to vary between tasks since encoding times of different types of information -related to each task- vary (reflected on the distribution of minimum RT by task); inter-trial variability of the accumulation rate was included to account for slow errors (Ratcliff & McKoon, 2008). Crucially, we included by-participant random-intercepts over evidence threshold and accumulation rate, and random-slopes over non-decision time task effects.

We used the minimal model to build new models in which either the evidence threshold or the accumulation rate or both were allowed to vary by music condition since -*a priori*- we had no reasons to believe that the music effects were exclusive to a single DDM parameter. We

estimated these models by sampling 2000 traces with a 500 burn-in. Later, the winning model was better estimated with 12000 traces and a 5000 burn-in.

To diagnose estimation, we visually inspected MCMC chains and evaluated trace autocorrelation for each estimated coefficient.

To select the best model, we compared the Deviance Information Criterion (DIC) of the competing models as to select the one with the best trade-off between explanatory power and model complexity (i.e. lowest DIC), which yields similar results to the Watanabe-Akaike information criterion (WAIC) in evidence-accumulation models (Evans, 2019).

To validate the selected model, we performed PPC as for the GLMMs, only that here two conditional RT distributions are represented (for correct and error responses).

Model inference for hypothesis testing was based on analysis of conditional posterior distributions, as described in the main text of the article.

Models

Here, we report specification, priors, and validation for all models referenced in the main and supporting text and on which we based all inferences. For clarity, specifications are written with lme4-like syntax (Bates et al., 2015).

A. Decision outcomes

a. RT-GLMM

i. Specification.

1. Family:

Shifted log-Normal: $RT \sim e^{Normal(\mu,\sigma)} + ndt$

2. Formula:

μ~ task * (music * difficulty + resploc + trial) +

(task + music + difficulty + resploc + trial | pID)

log**(σ)** ~ task * (music + difficulty)

log**(ndt)** ~ task

3. Priors:

μ intercept ~ normal (-0.4,0.4)

log**(σ)** intercept ~ normal (-0.7,0.2)

log**(ndt)** intercept ~ normal (-1.7,0.1)

μ betas ~ normal (0,0.4)

σ betas ~ normal (0,0.2)

ndt betas ~ normal (0,0.1)

sd ~ normal (0.2,0.2)

ii. Validation.

1. $\overline{R^2} = 0.41$, $CI_{95} = [0.40,0.41]$

2. Posterior predictive checks (PPC) (Figure. S6a)

b. Accuracy-GLMM:

i. Specification.

1. Family:

Bernoulli: $Accuracy \sim Bernoulli(\pi)$

2. Formula:

logit**(π)** ~ task * (music * difficulty + resploc + trial) +

(task + music + difficulty + resploc + trial| pID)

3. Priors:

Intercept ~ student (3,0,2.5)

betas ~ normal (0,1.5)

sd ~ student (3,0,2.5)

ii. Validation.

1. Area under the Receiver Operating Characteristic curve: 0.79

2. PPC (Figure. S6b)

B. Subjective arousal

a. Arousal-rating ZOIB (music-elicited arousal)

i. Specification.

1. Family:

Zero-one-inflated-beta

$$Arousal \sim (1 - Bernoulli(\alpha)) * Beta(\phi * \mu, \phi * (1 - \mu)) + Bernoulli(\alpha) * Bernoulli(\gamma)$$

2. Formula:

logit **(μ)** ~ music + (1 | pID)

log (**ϕ**) ~ 1

logit (**α**) ~ 1

logit (**γ**) ~ 1

3. Priors:

μ intercept ~ normal (0, 1.5)

μ betas ~ normal (0, 1.5)

ϕ intercept ~ student (3,0,2.5)

α ~ logistic (0,1)

γ ~ logistic (0,1)

sd ~ student (3, 0, 2.5)

ii. Validation.

1. $\overline{R^2}$ = 0.43, CI_{95} = [0.24,0.60]

2. PPC (Figure. S7a)

b. Arousal ZOIB (music-elicited arousal *during* decision-making)

i. Specification.

1. Family:

Zero-one-inflated-beta

$$Arousal \sim (1 - Bernoulli(\alpha)) * Beta(\phi * \mu, \phi * (1 - \mu)) + Bernoulli(\alpha) * Bernoulli(\gamma)$$

2. Formula:

logit **(μ)** ~ task * (music + block) +

(task * (music + block) | pID)

log (**ϕ**) ~ 1

logit (**α**) ~ 1

logit (**γ**) ~ 1

3. Priors:

μ intercept ~ student (3,0,2.5)

μ betas ~ normal (0, 1.5)

ϕ intercept ~ student (3,0,2.5)

α intercept ~ logistic (0,1)

γ intercept ~ logistic (0,1)

sd ~ student (3, 0, 2.5)

ii. Validation.

1. $\overline{R^2} = 0.85$, $CI_{95} = [0.84,0.85]$

2. PPC (Figure. S7b)

c. Arousal-RT GLMM (Arousal effect on mean RT)

i. Specification.

1. Family:

Shifted log-Normal: $RT \sim e^{Normal(\mu,\sigma)} + ndt$

2. Formula:

μ ~ task * (arousal + difficulty + block) +

(task + arousal + difficulty + block | pID)

log**(σ)** ~ task * (arousal + difficulty)

log**(ndt)** ~ task * arousal

3. Priors:

μ intercept ~ normal (-0.4,0.4)

log(σ) intercept ~ normal (-0.7,0.2)

log(ndt) intercept ~ normal (-1.7,0.1)

μ betas ~ normal (0,0.4)

σ betas ~ normal (0,0.2)

ndt betas ~ normal (0,0.1)

sd ~ normal (0.2,0.2)

ii. Validation.

1. $\overline{R^2} = 0.70$, $CI_{95} = [0.68,0.71]$
2. PPC (Figure. S7c)

C. Decision process (DDM)

a. Specification.

i. Family

WFPT (Navarro & Fuss, 2009):

$$(RT, Accuracy) \sim WFPT(v, a, z, t, sv)$$

ii. Formula

- Evidence-accumulation rate **(*v*)** ~ task * (difficulty + trial) + music
- Evidence threshold **(*a*)** ~ task * (music + trial)
- Non-decision time **(*t*)** ~ task + trial
- Starting point **(*z*)** = 0.5 (fixed)
- Accumulation rate inter-trial variability **(*sv*)** ~ task

iii. Priors (Wiecki et al., 2013). Note, HN: half-normal.

- ***v***: μ ~ Normal (2,3), σ ~ HN (2)
- ***a***: μ ~ Gamma (1.5,0.75), σ ~ HN (0.1)
- ***t***: μ ~ Gamma (0.4,0.2), σ ~ HN (1)
- ***sv***: μ ~ HN (2)

b. Selection

i. Deviance information criterion (DIC) comparison (Table S3)

c. Validation

i. PPC (Figure. S5)

Note, *music*: music condition (silence, slow -at 40bpm-, fast -at 190bpm-); *task*: decision task (RDK, Marble, City, Lexical, Snack); *difficulty*: trial difficulty (easy, medium, hard); *resploc:* response location (left, right); *trial*: scaled and centered trial number within a task; *pID*: participant ID; *mRT:* mean RT; *ndt*: non-decision time (shift); DDM: Drift-Diffusion Model; WFPT: Wiener first-passage time distribution; *betas*: prior for linear predictor coefficients.

Supplemental Figures and Tables

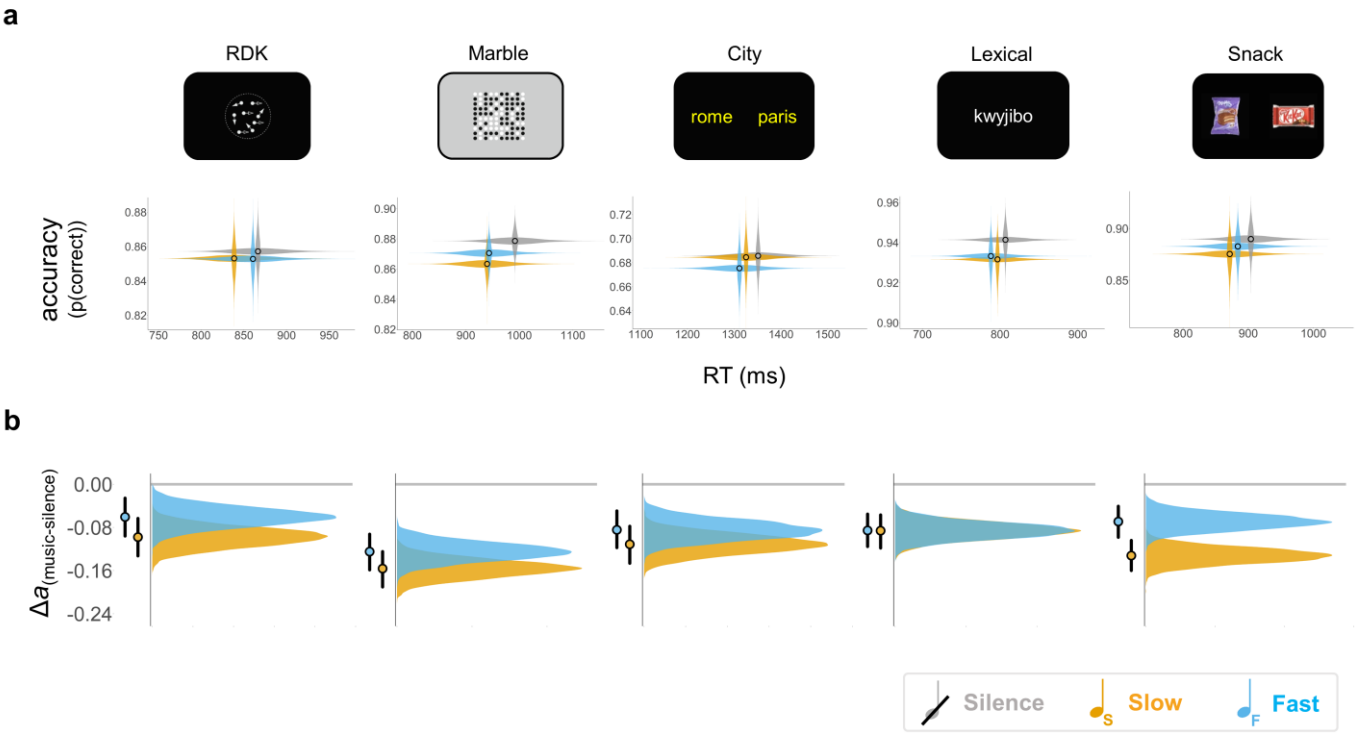

Figure S1. Music effects on decision-making by task.

a. Estimated RT and accuracy (i.e., probability of a correct response) means (averaged across difficulty and response location) for each task and music condition are represented jointly. Note that, under the Bayesian framework, model estimation returns a probability distribution of values for each parameter that is the result of combining the prior probability for each parameter and the likelihood of the observed data given a parameter value, called the updated or posterior distribution (henceforth, posterior). Here, estimated-means posteriors are represented as "eye" intervals -with higher probability mapped to greater width- and the point-estimates (the posterior median) are represented by colored dots. Contrasts between conditions (Table 1) were performed on the linear predictor scale (log and logit, for RT and accuracy, respectively), but we transformed these estimates (to linear time units and odds ratio, respectively) to aid interpretability. These effects were traced to **b.** A reduction in evidence threshold (*a* in DDM). To mimic the DDM graphical depiction, coordinates were flipped. On the vertical axis we represent

the change in threshold relative to silence condition (grey horizontal line). The horizontal axis represents the probability density of the contrast values. The black vertical segments represent the CI_{95} and the embedded colored dots are the medians of the contrast distributions. Both slow and fast music significantly lowered the decision threshold relative to silence on all tasks (Table 1). All results are based on data from the 32 participants.

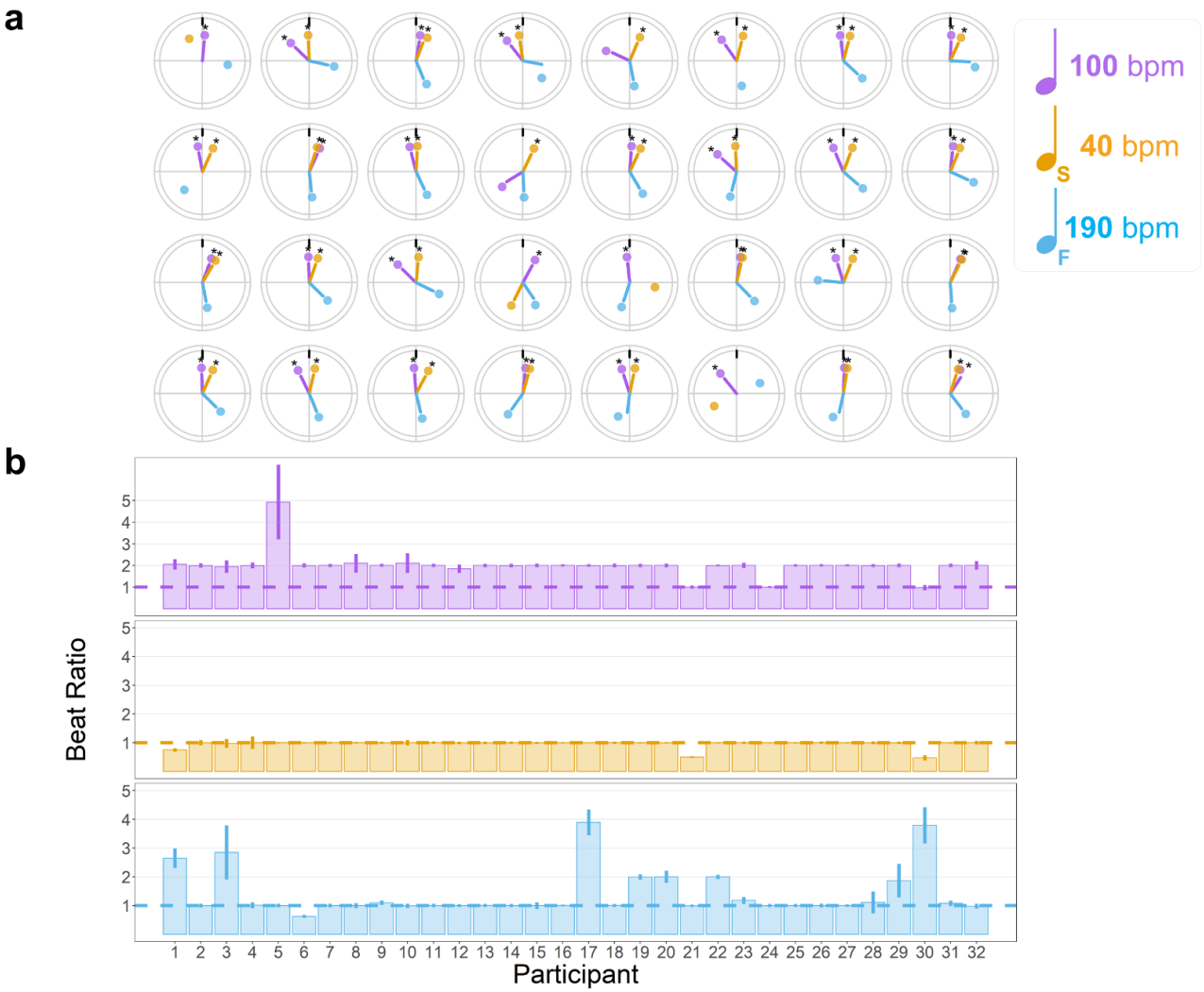

Figure S2. Participants were able to tap to the *tempo* (beat) of both slow and fast music.

We used circular statistics (Fitak & Johnsen, 2017; Landler et al., 2018) to assess whether participants were able to tap to the beat of unrelated (training, purple) music, and of slow and fast music used in the main decision-making experiment. **a.** After transforming keypress-data to circular data (relative to the beat of the music that was being played), we assessed unimodality with the Rayleigh test. Clock-plots were constructed with tapping data from each of the 32 participants (from left to right, and from top to bottom) and when unimodality was significant (represented by the presence of a solid radial line), the estimated circular mean (the angle of the solid radial lines) generally matched the observed mean (colored dot). For each clock-plot,

the small vertical black line indicates the start of the beat cycle (0), with time following a clockwise direction. Asterisks denote when the estimated mean was not different from 0 (i.e., participant was tapping on the beat) assessed with a directional V-test. If participants were tapping to a different (non-multiple) beat, taps would shift phase on every cycle, since each cycle is referenced to the true beat, which would be evidenced as a departure from unimodality. Further, tapping exactly to the beat of the fast music was very challenging since each beat duration is 60/190=0.31s. So, phase-shifts were expected. With very few exceptions, participants were generally able to perceive and tap to the beat (92.7% of tapping data). Phase-synchronization was less common (60,4%), mainly because of the shift on fast music. In any case, beat synchronization (and not phase) was the relevant feature to our hypotheses. However, these results do not answer whether participants were tapping exactly at every beat or rather at multiples (e.g., every 2 or 3 beats), which would be relevant to determine whether participants accurately perceived the actual tempo of the music (i.e., if they tapped at every 2 beats for the fast music, the subjective tempo would be half (95bpm) the actual tempo (190bpm)). **b.** Although for training music (at 100bpm, purple) most participants tapped every 2 beats, subjective beat matched the actual beat (i.e., participants did not tap over multiples of the beat) for slow and fast music, which is depicted with dashed lines. Colored bars and solid vertical lines represent mean beat ratio ± SD for each participant.

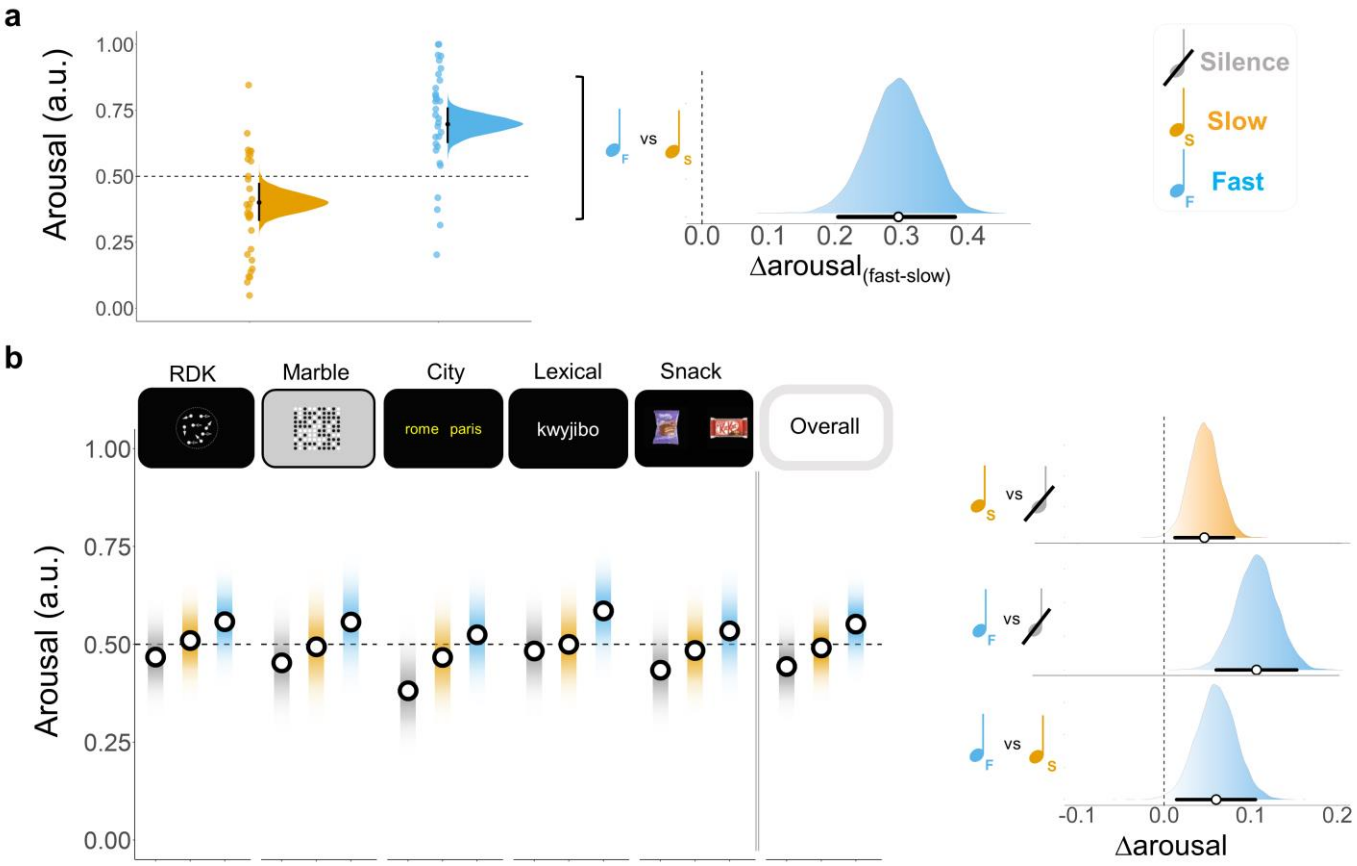

Figure S3. Faster music *tempo* induces higher subjective arousal levels.

Fast music induced higher arousal levels both after and during decision-making. **a.** After listening to a 12-second excerpt of each musical track (the same we used as our experimental conditions), participants rated their subjective arousal with a visual analogue scale (ranging from 0: *low* to 1: *a lot*). We represent the estimated means on the left-side plot in which scattered dots are the raw data (jittered on the x axis for clarity), white dots are the medians of the posterior distributions (colored densities) for each music and black vertical lines represent CI_{95} of the posteriors. Hypothesis testing was performed over the difference between conditions for arousal ratings (right-side plot), revealing that fast music elicited higher arousal than slow music. **b.** The left-side plot shows the estimated posterior medians (white points) and their probability

(mapped to color opacity) for each task and for the task average by music condition. On the right side, posterior density plots (with greater effects -i.e., greater differences- mapped to higher density color opacity) for the difference between music conditions (overall) revealed increasingly higher arousal levels with slow and fast music, in that order.

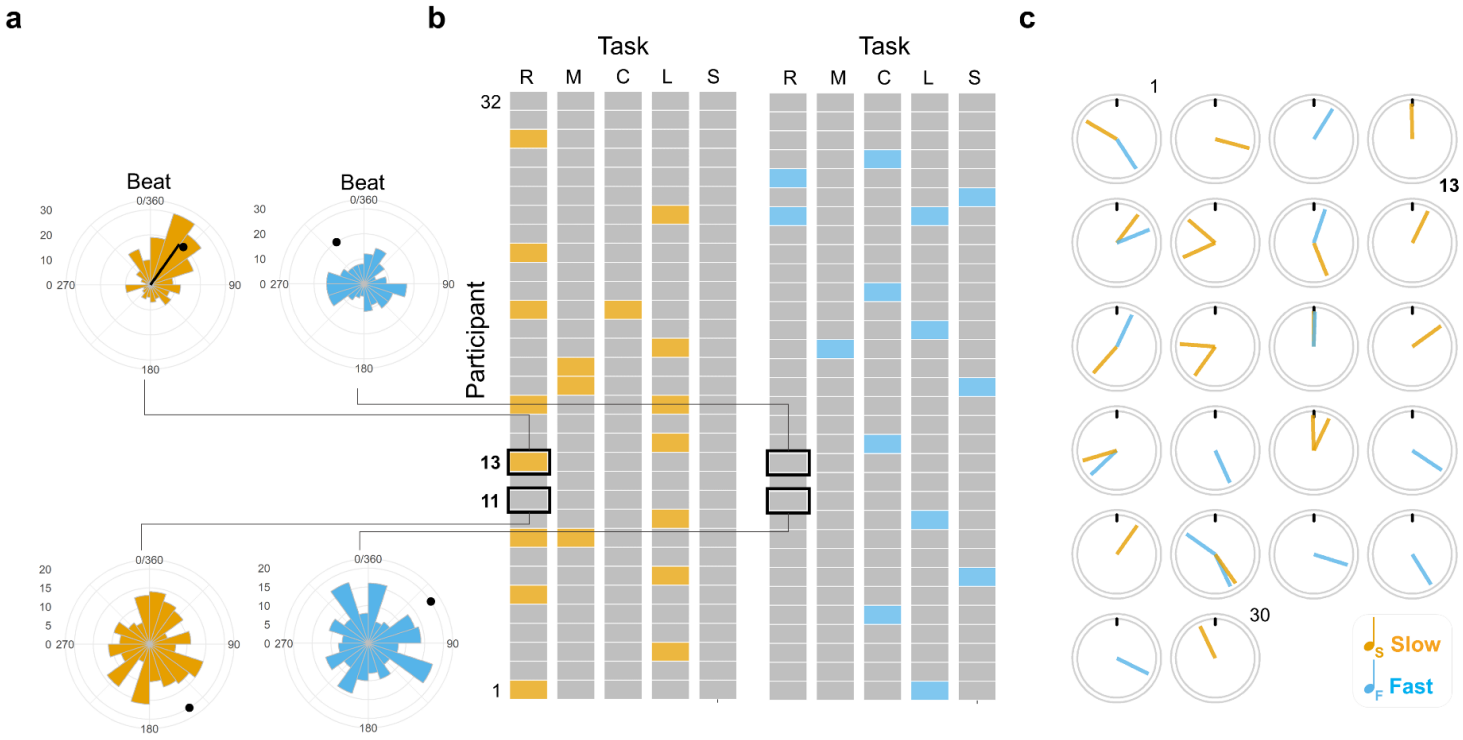

Figure S4. Response times are not a result of a synchronize-to-the-beat strategy.

If participants responded at the expected beat of the background music, histograms of the differences between keypress times and the actual beat-time should reveal a unimodal distribution. If synchronization was perfectly accurate, the value of this mode (phase) should be 0. **a.** Plots are circular histograms for the transformed keypress-time data for a given participant, task (RDK), and music condition. Black dots represent the observed circular means. Values on the circles' perimeter are angles –phase shifts- relative to the beat (at 0/360°, vertical upwards), *e.g.,* a slow-music data-point at 180° means that the keypress for that trial was made 0.75s before/after the actual music beat. Rayleigh-test for unimodality was performed to assess whether the estimated mode was statistically significant. For clarity, we show data for two participants with different behavior. On the lower margin, circular plots for responses of participant 11 (in the RDK task for slow and fast music) are an example of non-synchronized behavior and are representative of most participants. Note, although a circular mean can be

calculated (black dot) it is non-informative as data show no evidence of departure from uniformity. On the upper margin, circular plots for responses of participant 13 in the RDK task for slow music show that responses followed a consistent period (unimodality) and its phase is represented by a solid black line. **b.** Participant-by-task "punch-card" plot (for each music condition) show when responses from a given participant at a given task (R: RDK, M: Marble, C: City, L: Lexical, S: Snack) were unimodal (i.e., followed the beat). Only 10% of the data had significant unimodal distributions and were not consistent either between or within participant (e.g., a participant synchronized on RDK trials with slow music, but not for fast in the same task). **c,** If a participant with unimodal distribution (each clock-plot) effectively synchronized, we expected that phases (arrow direction) would be consistent within that participant for different tasks (but not necessarily for different *tempi*), however we did not observe this pattern. The beat is denoted by the vertical black line (at "12 o'clock" of each clock-plot). Overall, these results show that response times were likely not a product of a synchronization strategy.

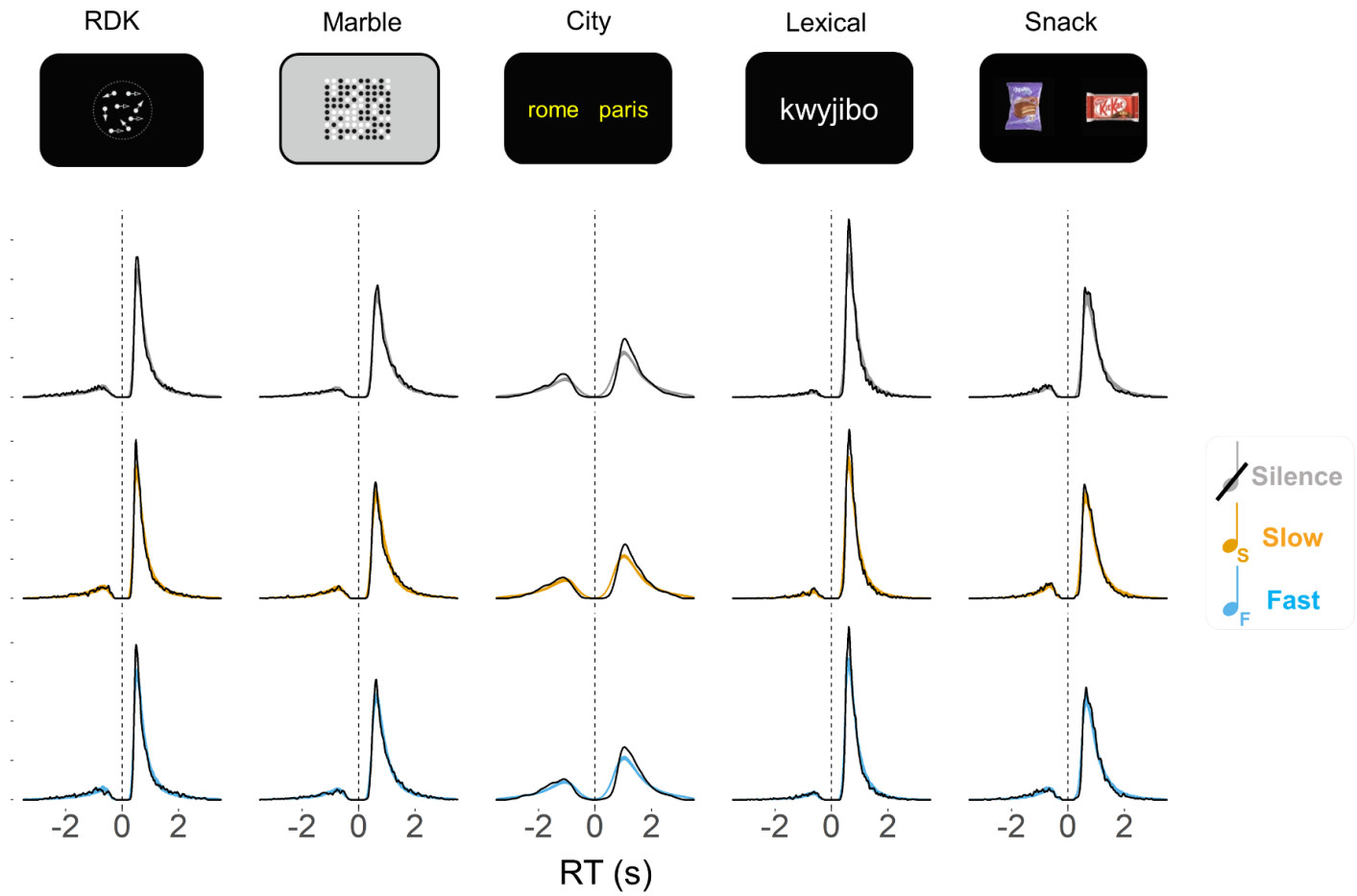

Figure S5. Posterior predictive checks (PPC) for the decision process model (DDM).

To validate the winning DDM model (for the effect of music on the decision process) we performed PPC as described in the Methods section. Colored lines represent distributions of 100 randomly selected datasets generated from the posterior distribution of the model. Black lines are the distribution of observed data. Rows divide music conditions (from top to bottom: silence, slow, fast) and columns divide tasks (from left to right: RDK, Marble, City, Lexical, Snack). Vertical axis represents probability density, and the horizontal axis represents RT (note, RT for correct (error) responses are represented with positive (negative) values). The relative height of the negative and positive densities (divided by the vertical dashed line) in each panel reflects the accuracy for that condition. The fit was generally good, although for the City task there were distributional features mismatches.

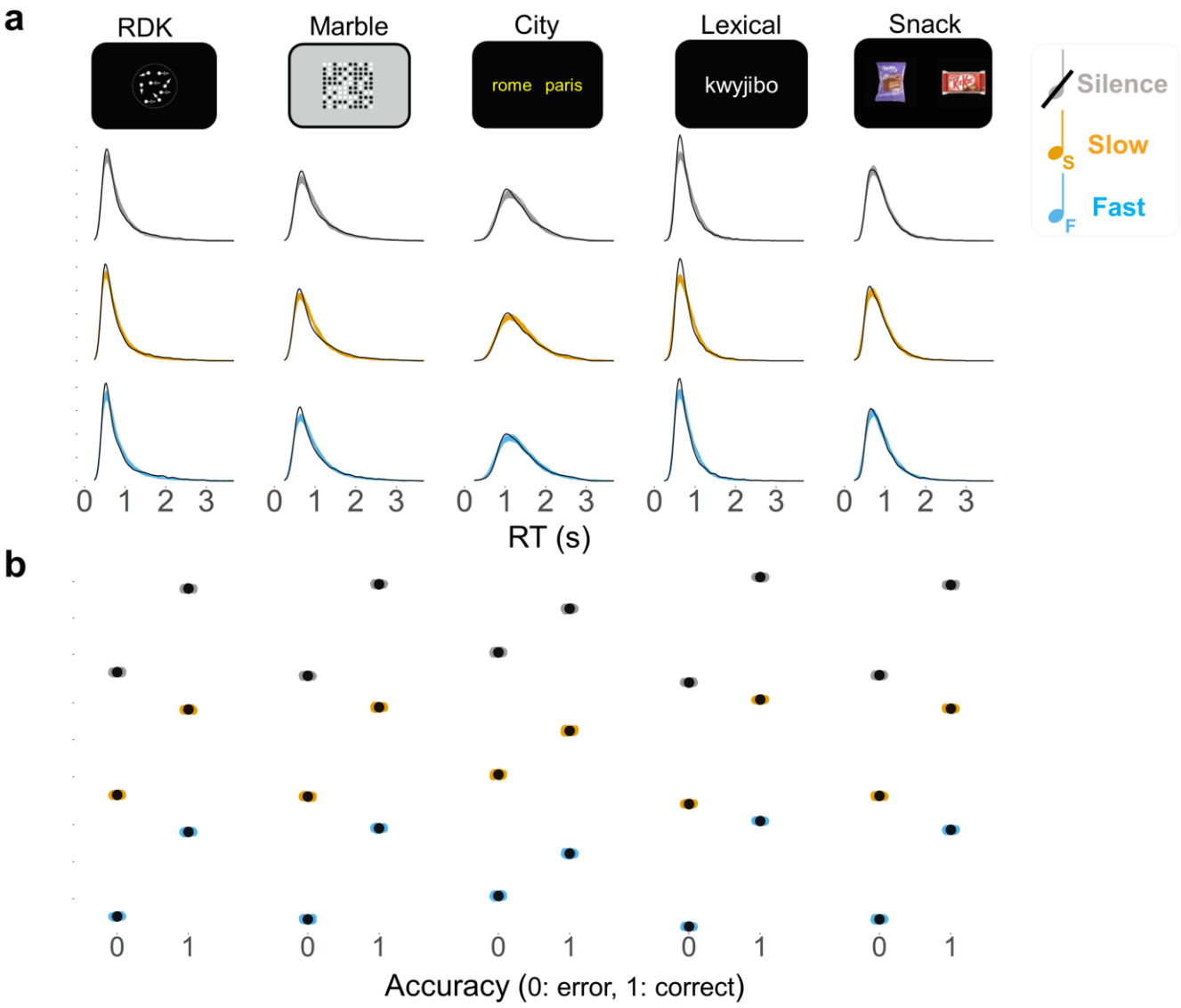

Figure S6. PPC for decision outcomes (RT and accuracy).

To validate the RT and accuracy models we performed PPC as described in the Methods section. Rows divide music conditions (from top to bottom: silence, slow, fast) and columns divide tasks (from left to right: RDK, Marble, City, Lexical, Snack). Vertical axis represents probability, and the horizontal axis represents the dependent variable. **a.** For the RT-GLMM (effect of music on RT), colored lines represent distributions of 100 randomly selected datasets generated from the posterior distribution of each model. Black lines are the distribution of observed data. Fit was generally good, although for Marble and Lexical, some distributional

features are slightly off (which is consistent with the point-wise estimation of R^2=0.41). **b.** For the accuracy-GLMM (effect of music on decision accuracy), posterior predictions (on the response variable scale, i.e. 0: error, 1: correct response) from 100 random samples (colored dots, jittered on the x axis for clarity) are plotted against the actual accuracy data (black dots). Fit was generally good, as predictions are virtually undistinguishable from actual data, which is in line with the estimated AUC-ROC of 0.79).

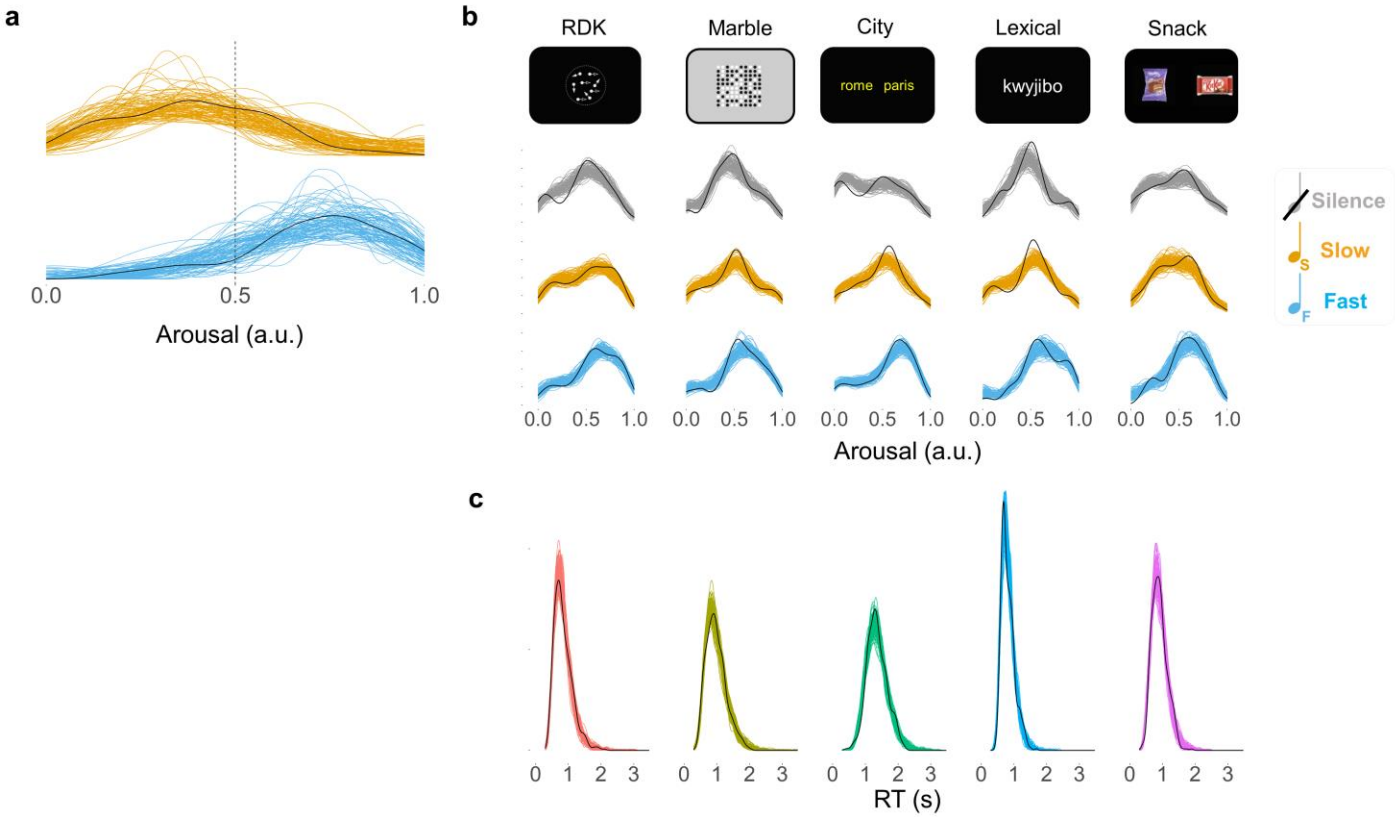

Figure S7. PPC for arousal models.

To validate the arousal-related models we performed PPC as described in the Methods section. Colored lines represent distributions of 100 randomly selected datasets generated from the posterior distribution of each model. Black lines are the distribution of observed data. **a.** For the Arousal-rating ZOIB (effect of music on arousal ratings) rows divide music conditions (slow and fast). The fit captured the main distributional features although with a somewhat wide uncertainty (which is consistent with the point-wise estimation of R^2=0.43). **b.** For the Arousal-ZOIB (effect of music on arousal during decision-making) rows divide music conditions (silence, slow and fast) and columns divide tasks (from left to right: RDK, Marble, City, Lexical, Snack). The fit was generally good which is in line with R^2=0.85. **c.** For the Arousal-RT GLMM (arousal effect on mean RT) columns divide tasks (from left to right: RDK, Marble, City, Lexical, Snack).

The fit was generally good which is in line with R^2=0.70. In all cases, the vertical axis represents probability density, and the horizontal axis represents the dependent variable.

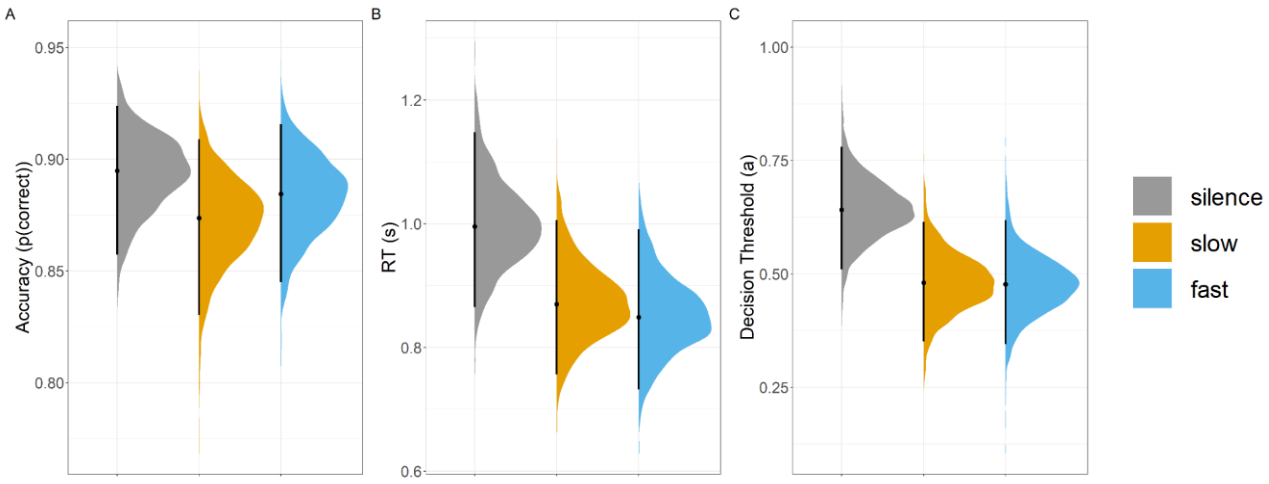

Figure S8. Effects of music at a less-disruptive slow tempo (70 bpm) on decision-making.

Estimated (A) accuracy, (B) RT and (C) decision threshold (*a* parameter in DDM) conditional means ((and their corresponding 95% CI) for the silence (gray), new slow (70bpm, orange) and fast (190bpm, light blue) music conditions. All music effects are reliable (i.e., the 95% CI of the contrasts' posterior distribution do not include zero), except for the odds ratio estimates (for assessing effects on accuracy) which, although have most of their posterior density below zero, these posteriors include zero within their 95% CI.

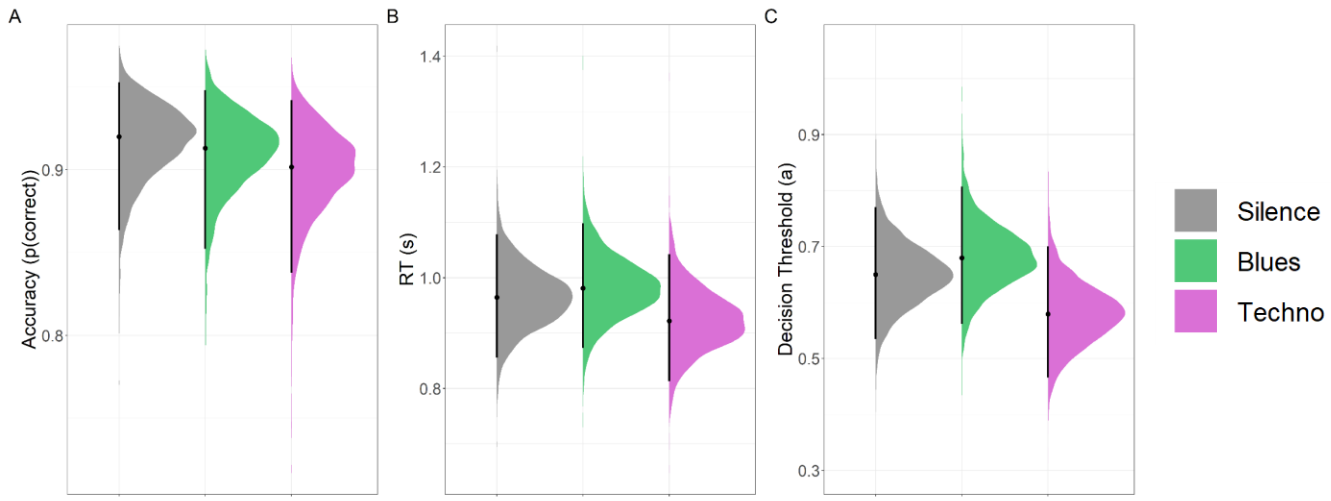

Figure S9. Effects of blues and techno music on decision-making.

Estimated (A) accuracy, (B) RT and (C) decision threshold (*a* parameter in DDM) conditional means ((and their corresponding 95% CI) for the silence (gray), *blues* (green) and *techno* (purple) music conditions (both at 190 bpm). Blues music had minimal impact on decisions (slightly slower yet inaccurate decisions, mapping to a marginally more cautious decision threshold). Techno music produced similar effects as those described in the original manuscript. Note that all the techno music effects are reliable (i.e., the 95% CI of the contrasts' posterior distribution do not include zero), except for the odds ratio estimate (for assessing effects on accuracy) which, although has most of their posterior density below zero, these posteriors include zero within their 95% CI.

Table S1. Estimated effects of difficulty over music condition effects (i.e., interaction effects) on RT and Accuracy.

difficulty.effect	music.effect	median	.lower	.upper	dep.var
Hard-Easy	Slow-Silence	-18.51	-35.14	-2.41	RT
Hard-Medium	Slow-Silence	-16.65	-33.92	0.40	RT
Medium-Easy	Slow-Silence	-1.99	-14.29	10.31	RT
Hard-Easy	Fast-Silence	-8.37	-25.57	8.07	RT
Hard-Medium	Fast-Silence	-10.98	-28.51	5.87	RT
Medium-Easy	Fast-Silence	2.71	-9.77	15.17	RT
Hard-Easy	Fast-Slow	10.19	-6.40	26.49	RT
Hard-Medium	Fast-Slow	5.63	-11.63	22.48	RT
Medium-Easy	Fast-Slow	4.65	-7.61	16.91	RT
Hard-Easy	Slow-Silence	0.13	-0.03	0.28	Accuracy
Hard-Medium	Slow-Silence	0.09	-0.02	0.20	Accuracy
Medium-Easy	Slow-Silence	0.03	-0.13	0.20	Accuracy
Hard-Easy	Fast-Silence	0.15	-0.01	0.30	Accuracy
Hard-Medium	Fast-Silence	0.05	-0.06	0.16	Accuracy
Medium-Easy	Fast-Silence	0.10	-0.07	0.26	Accuracy
Hard-Easy	Fast-Slow	0.02	-0.13	0.17	Accuracy
Hard-Medium	Fast-Slow	-0.04	-0.15	0.07	Accuracy
Medium-Easy	Fast-Slow	0.06	-0.10	0.22	Accuracy

Primary results indicated are RT and accuracy (*dep.var*) median values (*median*) of the posterior difference between difficulties (*difficulty.effect*) for each music condition contrast (*music.effect*) and the corresponding lower (.*lower*) and upper (.*upper*) 95% credible interval bounds.

Table S2. Estimated conditional (by task) and marginal (across tasks) coefficients for the effect of arousal on decision RT and accuracy.

task	beta.arousal	lower.HPD	upper.HPD	dep.var
RDK	-0.1	-0.25	0.063	RT
MRB	0.089	-0.085	0.267	RT
CTY	-0.088	-0.191	0.027	RT
LEX	-0.119	-0.321	0.055	RT
SNK	0.03	-0.11	0.178	RT
overall	-0.038	-0.13	0.053	RT
RDK	0.14	-0.294	0.589	Accuracy
MRB	-0.028	-0.647	0.628	Accuracy
CTY	-0.282	-0.804	0.281	Accuracy
LEX	0.209	-0.582	0.989	Accuracy
SNK	-0.062	-0.826	0.685	Accuracy
overall	-0.008	-0.33	0.31	Accuracy

Primary results indicated are median values (*beta.arousal*) and the lower (*lower.HPD*) and upper (*upper.HPD*) 95% credible interval bounds.

Table S3. DDM selection by Deviance information criterion (DIC).

Model	v							a					t		sv	DIC
	Task	Difficulty	Music	Trial	Task: Difficulty	Task: Music	Task: Trial	Task	Music	Trial	Task: Music	Task: Trial	Task	Trial	Task	
1 (winning)	•	•	•	•	•		•	•	•	•	•	•	•	•	•	130017.25
2	•	•	•	•	•		•	•	•	•		•	•	•	•	130021.05
3	•	•		•	•		•	•	•	•	•	•	•	•	•	130027.05
4	•	•	•	•	•	•	•	•	•	•		•	•	•	•	130027.13
5	•	•		•	•		•	•	•	•		•	•	•	•	130031.24
6	•	•	•	•	•		•	•		•		•	•	•	•	130261.74
7	•	•		•	•		•	•		•		•	•	•	•	130263.08
8	•	•		•				•		•			•	•	(•)	136646.05
9																189001.16

●: predictor included in the model

(●): only intercept term included in the model

Trial: chronological predictor (trial number, centered and scaled)

***v*:** evidence-accumulation rate

***a*:** evidence threshold

t: non-decision time

***sv*:** evidence-accumulation rate inter-trial variability

In all cases, starting point (z) was fixed at middle-point (unbiased z = 0.5)

Note: Divided columns indicate interaction terms and colors are only used as visual aid to identify and distinguish between model parameters' linear dependencies.

Supplemental Material References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spat Vis*, *10*(4), 433-436. https://www.ncbi.nlm.nih.gov/pubmed/9176952

Brysbaert, M., & Stevens, M. (2018). Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *J Cogn*, *1*(1), 9. https://doi.org/10.5334/joc.10

Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Duchon, A., Perea, M., Sebastian-Galles, N., Marti, A., & Carreiras, M. (2013). EsPal: one-stop shopping for Spanish word properties. *Behav Res Methods*, *45*(4), 1246-1258. https://doi.org/10.3758/s13428-013-0326-1

Dutilh, G., & Rieskamp, J. (2016). Comparing perceptual and preferential decision making. *Psychon Bull Rev*, *23*(3), 723-737. https://doi.org/10.3758/s13423-015-0941-1

Evans, N. J. (2019). Assessing the practical differences between model selection methods in inferences about choice response time tasks. *Psychon Bull Rev*, *26*(4), 1070-1098. https://doi.org/10.3758/s13423-018-01563-9

Fitak, R. R., & Johnsen, S. (2017). Bringing the analysis of animal orientation data full circle: model-based approaches with maximum likelihood. *J Exp Biol*, *220*(Pt 21), 3878-3882. https://doi.org/10.1242/jeb.167056

Gonzalez-Nosti, M., Barbon, A., Rodriguez-Ferreiro, J., & Cuetos, F. (2014). Effects of the psycholinguistic variables on the lexical decision task in Spanish: a study with 2,765 words. *Behav Res Methods*, *46*(2), 517-525. https://doi.org/10.3758/s13428-013-0383-5

Goodrich, B., Gabry, J., Ali, I., & Brilleman, S. (2018). *rstanarm: Bayesian applied regression modeling via Stan. R package version 2.17. 4*. In http://mc-stan.org

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *J Methods in Ecology Evolution*, *7*(4), 493-498.

Husain, G., Thompson, W. F., & Schellenberg, E. G. (2002). Effects of Musical Tempo and Mode on Arousal, Mood, and Spatial Abilities. *Music Perception*, *20*(2), 151-171. https://doi.org/10.1525/mp.2002.20.2.151

John Iversen, A. P. (2008). The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population. *Proceedings of the 10th International Conference on Music Perception and Cognition*, 465-468.

Kay, M. (2020). *tidybayes: Tidy data and geoms for Bayesian models*. In (Version 2.0.3)

Keuleers, E., & Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. *Behav Res Methods*, *42*(3), 627-633. https://doi.org/10.3758/BRM.42.3.627

Landler, L., Ruxton, G. D., & Malkemper, E. P. (2018). Circular data in biology: advice for effectively implementing statistical procedures. *Behav Ecol Sociobiol*, *72*(8), 128. https://doi.org/10.1007/s00265-018-2538-y

Milosavljevic, M., Malmaud, J., Huth, A., Koch, C., & Rangel, A. (2010). The drift diffusion model can account for the accuracy and reaction time of value-based choices under high and low time pressure. *Judgment & Decision Making*, *5*(6), 437.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of mathematical psychology*, *53*(4), 222-230.

Ospina, R., & Ferrari, S. L. (2012). A general class of zero-or-one inflated beta regression models. *Computational Statistics & Data Analysis*, *56*(6), 1609-1623.

Palmer, J., Huk, A. C., & Shadlen, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J Vis*, *5*(5), 376-404. https://doi.org/10.1167/5.5.1

Parasuraman, R., & Mouloua, M. (1987). Interaction of signal discriminability and task type in vigilance decrement. *Perception & Psychophysics*

41(1), 17-22.

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. *Psychol Rev*, *117*(3), 864-901. https://doi.org/10.1037/a0019737

R Core Team. (2020). *R: A Language and Environment for Statistical Computing*. In https://www.R-project.org/

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision Sciences*, *2*(4), 237.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychol Rev*, *111*(1), 159-182. https://doi.org/10.1037/0033-295X.111.1.159

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput*, *20*(4), 873-922. https://doi.org/10.1162/neco.2008.12-06-420

Reinhardt, J. (2020). *bayesWilcoxTest: A Bayesian Alternative to the Wilcoxon Signed Rank Test*. In (Version 0.1.0)

Shadlen, M. N., & Kiani, R. (2013). Decision making as a window on cognition. *Neuron*, *80*(3), 791-806. https://doi.org/10.1016/j.neuron.2013.10.047

van den Brink, R. L., Murphy, P. R., & Nieuwenhuis, S. (2016). Pupil diameter tracks lapses of attention. *PloS one*, *11*(10), e0165274.

Wagenmakers, E. J., & Brown, S. (2007). On the linear relation between the mean and the standard deviation of a response time distribution. *Psychol Rev*, *114*(3), 830-841. https://doi.org/10.1037/0033-295X.114.3.830

Wearden, J. H., Norton, R., Martin, S., Montford-Bebb, O. J. J. o. E. P. H. P., & Performance. (2007). Internal clock processes and the filled-duration illusion. *33*(3), 716.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Front Neuroinform*, *7*, 14. https://doi.org/10.3389/fninf.2013.00014

Zylberberg, A., Barttfeld, P., & Sigman, M. (2012). The construction of confidence in a perceptual decision. *Front Integr Neurosci*, *6*, 79. https://doi.org/10.3389/fnint.2012.00079