Optimism, pessimism and judgement bias in animals: a systematic review and meta-analysis

Lagisz, Malgorzata [1†], Zidar, Josefina[2,†], Nakagawa, Shinichi[1,†*], Neville, Vikki[3], Sorato, Enrico[2], Paul, Elizabeth S.[3], Bateson, Melissa[4], Mendl, Michael[3#], Løvlie, Hanne[2#]

Addresses:

[1] Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Sydney, NSW 2052, Australia

[2] The Department of Physics, Chemistry and Biology, IFM Biology, Linköping University, SE-581 83 Linköping, Sweden

[3] Centre for Behavioural Biology, Bristol Veterinary School, University of Bristol, Langford, BS40 5DU, United Kingdom

[4] Centre for Behaviour and Evolution, Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom

***Correspondence**: Shinichi Nakagawa s.nakagawa@unsw.edu.au

[†] These authors contributed equally to this work

[#] These authors supervised this work equally and are joint senior authors

Author contributions: two groups of the authors, SN & HL and MM, EP & MB conceived the idea independently, SN developed study design and methods from the inputs from others. JZ, ES, VN and EP collected the data with inputs from SN, ML, MB, MM and HL. ML, JZ and SN conducted the analysis with inputs from MM. JZ, ML, SN, VN and MM co-wrote the first draft and all contributed to revisions of the manuscript.

Declarations of interest: none

Abstract

Just as happy people see the proverbial glass as half-full, 'optimistic' or 'pessimistic' responses to ambiguity might also reflect affective states in animals. Judgement bias tests, designed to measure these responses, are an increasingly popular way of assessing animal affect and there is now a substantial, but heterogeneous, literature on their use across different species, affect manipulations, and study designs. By conducting a systematic review and meta-analysis of 459 effect sizes from 71 studies of non-pharmacological affect manipulations on 22 non-human species, we show that animals in relatively better conditions, assumed to generate more positive affect, show more 'optimistic' judgements of ambiguity than those in relatively worse conditions. Overall effects are small when considering responses to all cues, but become more pronounced when non-ambiguous training cues are excluded from analyses or when focusing only on the most divergent responses between treatment groups. Task type (go/no-go; go/go active choice), training cue reinforcement (reward-punishment; reward-null; reward-reward) and sex of animals emerge as potential moderators of effect sizes in judgement bias tests.

Keywords: research synthesis, affective state, cognitive bias, animal welfare

Introduction

Accurate assessment of affect (emotion) in non-human animals is an important goal in disciplines including animal welfare science, neuroscience, psychopharmacology and drug development. A prevailing view in the study of human emotion is that affective states comprise subjective, behavioural, neural and physiological components (Paul et al., 2020; Scherer, 2005). Whilst the subjective component of animal affective states (feelings) is not currently accessible to direct measurement and we cannot be certain which species consciously experience such states (see Paul et al., 2020), we can objectively assess the other components. In his book *The Expression of Emotions in Man and Animals*, Darwin (1872) focused on behavioural manifestations of animal emotion, namely "expressive movements of the face and body"", and such measures continue to be used as indicators of animal affect today (e.g. Girard & Bellone, 2020). But other measures focus more directly on the role of affect in behavioural control and decision-making. A relatively new and promising approach is to measure biases in decision-making under ambiguity as indicators of animal affect (Harding et al., 2004; Mendl et al., 2009). This is because there are empirical and theoretical reasons to expect that responses to such ambiguity reflect affective valence (positivity or negativity of an affective state). For example, people in negative states are more likely to make negative ('pessimistic') judgements about ambiguous events or stimuli than people in more positive states (Blanchette and Richards, 2010; Paul et al., 2005). Such assessments could reflect an adaptive use of background affect (or mood) as a Bayesian prior over the likelihood of future positive or negative outcomes (Mendl et al., 2010; see Mendl & Paul, 2020 for a fuller discussion).

In line with these findings and ideas, a generic assay for measuring these so-called '*judgement biases*' has been developed for animals and has now been used in a large number of studies across a range of species. The original assay (Harding et al., 2004) involves training subjects to make one response (positive response) to a 'positive' cue (a single frequency tone) in order to achieve a positive outcome (e.g. food) and a different response (negative response) to a 'negative' cue (a tone of a different frequency) in order to avoid a negative outcome (e.g. white

noise) (Figure 1a). Once subjects have learnt this conditional discrimination task, training continues but includes occasional ambiguous cues (tones of intermediate frequency) designed to assess whether subjects would make the positive response indicating anticipation of a positive outcome, or the negative response indicating anticipation of a negative outcome. This allows one to test whether, for example, animals in a putative negative affective state (e.g. as a result of some sort of experimental treatment, Figure 1b) are more likely to make the negative response, as predicted (Figure 1c,d). Making the positive or negative response under ambiguity can be operationally defined as 'optimistic' or 'pessimistic' (Bateson, 2016) without implying that animals experience optimism or pessimism as humans do.

Published studies using this judgement bias task (also referred to as an 'ambiguous cue interpretation' task (Rygula et al., 2013)), and variants of it, have supported the general prediction, but also generated null and opposite results. These findings have been summarised narratively in a number of review papers that have also identified various methodological and theoretical questions regarding the task and approach (Baciadonna and McElligott, 2015; Bethell, 2015; Gygax, 2014; Hales et al., 2014; Mendl et al., 2009; Mendl & Paul 2020; Roelofs et al., 2016). What has been lacking, and much needed, is a systematic review and meta-analysis of the findings to date to evaluate whether the general predictions behind the approach are supported, and how results may be influenced by a variety of moderators, including aspects of task design, methods used to manipulate affect, species studied, and age and sex of subjects. Recently, we published the first such meta-analysis focusing on the effects of pharmacological manipulations of affective state on judgement biases (Neville et al., 2020). Here we systematically review and meta-analyse the much larger number of studies that have used non-pharmacological affect manipulations.

We focus on judgement bias tasks based on the Harding et al. (2004) method, since these have been more widely studied in animals than other cognitive biases such as attention (Bethell et al., 2016; Crump et al., 2018) and memory biases (Burman and Mendl, 2018). Although details of

the procedures and criteria used to select suitable studies and extract appropriate data for the meta-analysis are explained in the Methods section, three points should be noted here.

First, a major challenge in any study of animal affect is to establish a 'ground truth' for the affective state that the animal is in when under study. This is necessary, for example, if an aim of a study is to determine what behavioural, physiological or neural changes occur in animals in particular affective states, and hence to develop reliable indicators of such states. Therefore, in studies which seek to evaluate whether judgement bias is a valid indicator of affective valence, we need to know whether the animal is in a relatively positive or negative state, so that we can test whether animals that are in a more positive state do indeed show more optimistic decisions under ambiguity, than those in a more negative state. In most judgement bias studies, researchers attempt to use an experimental treatment to induce a relatively positive or negative affective state compared to a control or 'benign' treatment group, or they impose both a positive and a negative treatment, and compare these. Because we cannot know for certain where the intermediate 'neutral' state lies, we use terminology that emphasises the relative nature of these manipulations. Thus, we refer to 'better' (more positive), 'benign/control', and 'worse' (more negative) treatments, and assign them to either 'relatively better' or 'relatively worse' groups for pair-wise comparison in the meta-analysis.

Second, there are two main types of task used in judgement bias trials: active choice (go/go) and go/no-go (Figure 1a). In go/go active choice tasks, the animal has to choose between two alternative responses (e.g. press the left or right lever), while in go/no-go tasks the animal's options are to perform a response (e.g. approach a location or press a lever) or supress it. The response of animals can be reported as a proportion (e.g. proportion of trials in which the subject pressed the left lever or approached the location), or a latency (e.g. time taken to press the lever or approach the location). Latency and proportion data have different statistical distributions; they require different transformations and use of different formulae to calculate effect sizes. There are also biological reasons for separating latency and proportion data.

Different measures may represent different aspects of cognitive processes and their utility depends on the type of cognitive bias task used. In go/no-go tasks, latency to perform the response under ambiguity is a direct measure of judgement bias. For example, if the positive response is to approach the cue, then quick approach to an ambiguous cue indicates an 'optimistic' response. In contrast, go/go active choice tasks require responses to both cues (e.g. press left or right lever), meaning that the latency to perform whichever response the animal selects is more difficult to interpret in terms of 'optimism' or 'pessimism'. Rather, the proportion of positive or negative responses provides more definitive information about 'optimistic' or 'pessimistic' decisions. This measure is also of use in go/no-go tasks. Therefore, proportion of positive vs. negative responses is preferable to latency as a measure of judgment bias for go/go active choice tasks, whereas for the go/no-go tasks both measures are, in principle, suitable.

Third, many judgement bias studies use more than one ambiguous cue during test trials (Figure 1c). Often three such cues are used; one (MID) which is assumed to be perceived by the animal as being at the mid-point of the sensory scale (e.g. sound frequency) between the positive (P) and negative (N) training cues, one (near positive: NP) which is half way between MID and the positive (P) cue, and one (near negative: NN) which is halfway between MID and the negative (N) cue. There are theoretical and methodological reasons for why an affect manipulation treatment might have an influence at one ambiguous cue but not at others in the same study. For example, non-midpoint ambiguous cues (NP, NN) may be perceptually too similar to the P and N training cues for animals to moderate their responses to them, whilst the midpoint (MID) cue is usually ambiguous enough for background affect to influence responses to it. In some studies MID could be perceived as closer to P or N and the most ambiguous cue becomes either NN or NP, respectively. Moreover, the perceived payoff of the positive and negative response outcomes, and hence associated decisions, may be asymmetrical. For example, if the perceived negative value of a foot-shock outcome is much stronger than the perceived positive value of a food pellet, animals may be strongly motivated to avoid shock risk and thus respond negatively to both MID and NN ambiguous cues, with variation in r response limited to the 'safest' NP

ambiguous cue (Mendl et al., 2009). Conversely, in a test variant where negative cues are simply lacking a reward instead of bearing a punishment, animals may respond positively even to negative cues, because the cost of doing so is negligible. Because it is likely that biased responses are unevenly spread across ambiguous cues – in fact some studies report effects which are strongest or only statistically significant at one ambiguous cue location (e.g. Bethell and Koyama, 2015; Zidar et al., 2018) – we investigate the effect of relative cue position and also conduct sensitivity analyses. These additional analyses use data subsets with different decision rules for selecting the most representative data points from response curves (e.g. using only the ambiguous cue with the largest absolute between-treatment effect size; more details in the Methods section).

This systematic review and meta-analysis aims to: (i) quantify the overall effect size that affect manipulations have on measures of judgement bias in animals; (ii) estimate heterogeneity of the results among different studies; (iii) explore the influences of different biological and methodological moderators (explanatory variables for variation in effect sizes).

Methods

Literature search

We conducted a systematic literature search and recorded relevant information required in the Preferred Reporting Items for Systematic reviews and Meta-Analyses statement (PRISMA; Moher et al., 2009; see Supplementary Materials for additional search details). We ran the first online database search on 29 October 2015, a second search in December 2017, and a final database search to update the dataset again on 27 March 2019. For these searches, we used the broad-coverage interdisciplinary databases *Scopus* and *Web of Science*, covering the titles, abstracts and keywords of academic publications.

The initial search string used in *Scopus* was: TITLE-ABS-KEY (("cognitive bias*" OR "judgment

bias*" OR "judgement bias*" OR "cognitive affective bias*") AND (pessimis* OR optimis* OR valence OR mood* OR emotion* OR "affective state*" OR "emotional state*" ambig* OR animal* OR "animal welfare")) AND PUBYEAR > 2003 and in *Web of Science*: TS=(("cognitive bias*" OR "judgment bias*" OR "judgement bias*" OR "cognitive affective bias*") AND (pessimis* OR optimis* OR valence OR mood* OR emotion* OR "affective state*" OR "emotional state*" ambig* OR animal* OR "animal welfare")) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article), Indexes=SCI-EXPANDED, SSCI Timespan=2004-2015. We restricted the publication years to those following the seminal paper on animal judgement bias (Harding et al., 2004). We restricted the subsequent updates of the literature search to the years since the previous search update (i.e. 2015-2017 and 2017-2019, respectively) and otherwise used the same search strings. We collected additional relevant studies from the authors whom we contacted to request data or other additional information that was missing from their publications. We also performed searches of reference lists of relevant review articles and research articles citing the seminal study by Harding et al. (2004).

The searches of the online databases generated over 900 potential article references and searches of other sources generated almost 500 additional references for screening (Figure 2). We removed duplicated results from these separate search paths. Two authors (J.Z. and E.S.) independently screened 482 abstracts from the articles identified in the 2015 search using the software AbstrackR (Wallace et al., 2012). M.L. performed two updates of literature searches in 2017 and 2019, following the same methodology as in the first search. Overall, we identified 74 published studies as potentially suitable for inclusion in our meta-analysis after screening of full texts and removal of duplicated studies. We excluded three studies during the data extraction stage (due to missing data), resulting in data from 71 studies being included in the meta-analysis.

Inclusion and exclusion criteria

We screened titles and abstracts from bibliometric records to identify empirical studies on

judgement bias in animals in which subjects were exposed to an affect manipulation aimed at inducing either a relatively positive or negative state. We then screened full text versions of the articles that passed this initial screening stage. At the full-text screening stage, the following six criteria had to be met for the study to be included in the meta-analysis: i) study had to be experimental and designed to investigate variation in judgement bias (i.e. 'optimistic' or 'pessimistic' interpretation of stimuli) in non-human animals; ii) experiments had to include at least two treatment groups (or control/'benign' and treatment groups); iii) experimental treatments had to be designed to induce 'relatively better' or 'relatively worse' affective states (see decision-tree in Supplementary Materials Figure S1); iv) for go/no-go tasks studies had to report either latency to make a response to ambiguous cues, or proportion of go or no-go responses towards ambiguous cues; for active choice tasks, studies had to report proportion of positive or negative responses; if the data available could be translated into such latencies or proportions, they were included; v) studies had to present data usable for effect size calculation; if suitable data could not be retrieved by contacting the authors, the study was excluded from the meta-analysis; vi) studies had to be published in peer-reviewed journals, but student reports and data from unpublished work, as well as articles that were written in languages other than English, could have been included if they met the above criteria.

We also excluded studies for the following additional reasons. We only considered data from studies investigating judgement bias, i.e. we excluded studies investigating other cognitive biases, such as attention bias and memory bias. We also excluded studies only describing judgement bias theory or methods or reviewing previous findings and studies that used the generic judgement bias task for humans, because our focus was on non-human animals. As studies investigating effects of drugs on judgement bias often include several doses that cannot easily be assigned into relatively better and relatively worse treatment groups, we also excluded all drug studies from this meta-analysis. As mentioned earlier, the drug studies were recently subjected to a separate meta-analysis by our group (Neville et al., 2020).

Data extraction

After compiling a final list of included studies, we extracted measurements representing behavioural responses to cues in the judgement bias tests. Each pairwise comparison consisted of a pair of outcome measures comparing behaviour of animals from 'relatively better' to 'relatively worse' affect manipulation groups. Our classification of treatments as inducing 'relatively better' or 'relatively worse' affective states was based on a decision tree involving screening articles and assessing treatments based on the following three criteria. First, if stated, we used the *a priori* hypothesis and reasoning outlined in the research article. Second, where possible, we employed Rolls' (2005, p.11) operational definition of emotion as "states elicited by rewards and punishers", where "a reward is anything for which an animal will work" and "a punisher is anything that an animal will work to escape or avoid". Thus, if a treatment involved stimuli that the subject animal is known to actively avoid, we deemed it to induce a relatively worse affective state than one which involved neutral or preferred stimuli. Third, we considered evidence from previous studies on the effects of the treatments in question on affective state (e.g. their effects on other putative indicators of affective state, such as abnormal repetitive behaviour or physiological stress indicators).

The decision tree for assigning affect treatments to relative affect manipulation categories is presented in Supplementary Materials Figure S1. If the first criterion in the decision tree was fulfilled (i.e. the authors of the original paper explicitly stated whether the treatment is expected to have positive/negative effect on animals' affective state), we ignored the subsequent decision criteria. If not, we evaluated the subsequent decision criteria. We classified all extracted treatment groups within a study relative to each other. For example, in a study with a control (benign/unmanipulated) and enriched housing group, the enriched group would be considered 'relatively better' and the control/benign group 'relatively worse'. Conversely, in a study with control/benign and stress-induction groups, the stress group would be considered 'relatively worse', and the control/benign group would be considered 'relatively better'.

We tackled variation in study design and outcome measurement as follows. First, for the go/no-go judgement bias tasks we extracted either or both (depending on which was reported) latency and proportion outcome measures (the signs of the effect sizes calculated from latency measurements were later inverted, so that interpretation of the effect direction was consistent with that for the proportion data). For active choice go/go judgement bias tasks, we extracted only proportion outcome measures (as explained earlier, latency measures in active choice tasks cannot be clearly linked to more 'optimistic' or 'pessimistic' responding). The extracted mean and standard error (or standard deviation) of responses to ambiguous and non-ambiguous cues during the tests were used to calculate values of effect sizes (and their variances) for each pairwise comparison of the relatively better and relatively worse treatment groups at the same cue. Relevant sample sizes were also recorded representing the number of animals from each group participating in the judgement bias test.

Second, included studies used varying numbers of ambiguous cues (range 1-13, mean 2.99, mode 3). We only extracted data for a maximum of three ambiguous cues per measurement (response curve). We always extracted data for the middle cue (midpoint between the positive and negative cues, MID) and, if available, two intermediate cues between the middle cue and positive and negative cues (near-positive NP and near-negative NN, respectively). If response data to positive (P) and negative (N) cues were reported for judgement bias tests, these were also extracted.

Third, when judgement bias was measured on several consecutive days following a treatment, we extracted the first measure only as it was usually closest in time to the acute affect manipulation treatment (Destrez et al., 2013; Doyle et al., 2011). In a few studies, animals were exposed to several judgement bias tests during a long-term treatment (Douglas et al., 2012; Hales et al., 2016; Rygula et al., 2013). In these cases, we extracted the last test occurring during each treatment, thus maximising the time available for it to exert its effects.

Fourth, some studies with a within-subject design measured judgement bias before, during and after an affect manipulation (e.g. pre-stress, stress, post-stress), or repeated the 'baseline' treatment (e.g. enriched, barren, enriched) (Brilot et al., 2010; da Cunha Nogueira et al., 2015; Hales et al., 2016; Murphy et al., 2013). In these studies, we compared measures taken before treatment ('baseline') to those taken during it and did not include measures taken after it.

Fifth, studies using a between-subject design sometimes tested both control/benign and treatment groups before, during and after a manipulation. In these cases, we compared the control/benign group to the treatment group during treatment and ignored the pre- and post-treatment measurements (Hales et al., 2016; Oliveira et al., 2016; Rygula et al., 2013).

Finally, if several treatments were applied where one or more treatments were hypothesized to be intermediate in effect to the two most extreme treatments, only the two extreme treatments were included (Ash and Buchanan-Smith, 2016; Burman et al., 2009; Keen et al., 2014; Wheeler et al., 2015).

For each experiment, we gathered information on the potential moderator variables to characterise our dataset and explain potential heterogeneity in the data. Detailed descriptions of all the originally extracted moderators are included in Table S1. In brief, the three key groups of extracted moderators considered information about the article, biological variables, and test design. Paper-specific information included authors, title, journal, and publication year. For each data point (i.e. comparison between two groups of animals), we extracted the following biological variables: taxa studied (mammals, birds, insects), sex (female, male, mixed-sex), age class (juvenile, adult) and source of animals (captive, wild-caught). Test-specific information included affect manipulation category (enrichment, stress, other), affect manipulation timing (before/during test, long-term), comparison category (Better-Worse, Benign-Worse, Better-Benign), type of cue used in judgement bias test (spatial, visual, auditory, tactile, olfactory), whether animals were food deprived prior to behavioural trials (yes, no/no information), automation of response measurement (yes, no/no information), blinding of personnel

performing trials (yes, no/no information), combination of reinforcement used during training (Reward Vs. Null, Reward Vs. Punishment; Reward Vs. Smaller Reward), task type (active choice go/go, go/no-go), whether ambiguous cues were reinforced (yes, no/no information), measurement type (latency, proportion), location of ambiguous cues relative to positive and negative cues (P – positive, NP – near-positive, MID – midpoint, NN – near-negative, N – negative). We also noted any pertinent additional details about study designs (between-subjects, within-subjects), affect manipulations, source of the data in the original studies, and any associated comments. When data were provided in a graph instead of a table or text, we extracted the values using GraphClick 3.0.3 (http://www.arizona-software.ch/graphclick/). Data extraction was performed by J.Z., M.L. and V.N. and was checked by M.L., V.N. and E.S.

Effect-size calculation

We used Hedges' unbiased standardized mean difference (Hedges' g) as the measure of effect size. Because latency and proportion data are bounded (i.e. latencies start at 0 and are often censored, and percentages are bounded between 0 and100, proportions between 0 and 1), we used natural log (for latencies) or logit-transformed data (for proportions, and percentages expressed as proportions) to calculate Hedges' g (details provided in Supplementary Materials Methods and Figure S2). In brief, to calculate Hedges' g, we focused on positive responses (i.e. those which indicated that the subject was anticipating a more rewarding outcome) and subtracted the mean value of the relatively worse treatment from the mean of the relatively better treatment, and divided the difference by the pooled standard deviation (SD) with correction for small sample sizes (Hedges and Olkin, 1985). Thus, if animals from the relatively better treatment group were making a higher proportion of positive responses than animals from the relatively worse treatment, the difference between the means would be positive and the effect size too. However, the expected pattern would be reversed when latencies to make the positive response were measured in go/no-go tasks: if animals from the relatively better treatment group were quicker to make the positive response (i.e. had lower latencies), than

animals from the relatively worse treatment, the difference between the means (and the resulting effect size) would be negative. To allow for easier comparison and interpretation of the effect sizes from latency and proportion measures, we reversed the sign of the effect sizes based on latency measures. Thus, after the sign adjustment, across all data positive values of Hedges' *g* can be interpreted as optimistic responses of animals exposed to relatively better treatments compared to those exposed to relatively worse treatments. For the go/no-go tests that reported the outcomes as both latency and proportion, we calculated Pearson's correlation between these two measures.

Meta-analysis and meta-regression models

We ran all statistical analyses in *R* version 3.6.0 (R Development Core Team, 2019); we created main forest-like (orchard) plots of effects using *orchaRd* package (Nakagawa et al., 2020). For multilevel meta-analysis and meta-regression we used the *rma.mv* function from the package *metafor* (Viechtbauer, 2010).

To estimate the overall mean of the effect sizes we constructed intercept-only models (i.e. meta-analysis) with study ID, experiment ID, cue ID, and effect size ID as random effects. To explore effect of species identity and phylogenetic relatedness, we also evaluated meta-analytic models with phylogeny and species ID added to the random effects list. We calculated I^2 values for each random factor and the overall heterogeneity, I^2_{Total}, in the meta-analytic models (Nakagawa and Santos, 2012).

To evaluate the effects of moderators of interest (e.g. subject sex or age class, test task type, test cue type and level of cue ambiguity), we ran univariate multilevel phylogenetic meta-regression models with moderators as fixed effects, and the same random effects as in the meta-analytic models (except species ID). In the multivariate meta-regression models (i.e. models with multiple moderators), we included only moderators that were significant in the univariate meta-regression models. We then performed AICc-based model selection using MuMIn package

(Barton, 2009) to infer relative contributions of included moderators. To assess the fit of meta-regression models, we calculated marginal R^2 values (*sensu* Nakagawa and Schielzeth, 2013; Nakagawa et al., 2017).

Publication bias

Statistically significant results are more likely to be published, resulting in a non-random sample of data available for meta-analysis (Rosenthal, 1979). To examine publication bias in our data set, we visually inspected a funnel plot for asymmetry in the distribution of the residuals of effect sizes (which are the sum of effect size level effects and sampling variance effects; i.e. meta-analytic residuals: *sensu* Nakagawa and Santos, 2012). We also performed Egger's regression on the residuals and measurement errors from the full meta-regression model (multilevel version of the publication bias test; Nakagawa and Santos, 2012). Egger's regression indicates publication bias if the regression intercept is significantly different from 0 (Egger et al., 1997). Finally, we tested for a special type of publication bias, a time-lag bias, i.e. a tendency for studies with larger effects to be published earlier (Jennions and Møller, 2002).

Sensitivity analyses (robustness of results)

To test robustness of our results to the estimation method, we ran a meta-regression model and a multilevel mixed-effect full meta-regression model (with subject sex, task type, cue type, and reinforcement type as moderators), using a Bayesian approach, as implemented in the *MCMCglmm* package (Hadfield, 2010). These models were run with 110,000 iterations, 10,000 burn-in periods, and thinning by every 100 resulting in an effective sample size of 1000. We used a parameter-expanded prior (V = 1, nu = 1, *alpha.mu* = 0, *alpha.V* = 1000), with EffectID (units) fixed at one.

We also ran the meta-analytic models using four additional data configurations representing different ways of interpreting results from pairs of response curves with multiple cues tested. First, we used a dataset with positive and negative test cues excluded, so that only responses to

ambiguous cues were used (maximum of 3 effect sizes per comparison of pair of response curves: for near-positive, midpoint, near-negative cues). In the remaining data subsets, we selected only one cue per response curve comparison. Thus, to create the second data subset, we only included data from the mid-point ambiguous cue location (MID data points and effect sizes). In the third data subset, we selected the effect sizes data from the cue location with the largest absolute value within each response curve comparison; notably, in 71.3% of the comparisons, the largest absolute effect size was not located at the mid-point ambiguous cue. In the fourth data subset, we used effect sizes with the biggest absolute value in the direction of the mean value, within each response curve comparison, as in Neville et al. (2020).

Results

Description of data set

The workflow and outcomes of our systematic literature searches are presented in a PRISMA diagram (Figure 2). The list of included studies is provided in Supplementary Table S2. Excluded studies, with reasons for exclusion, are listed in Supplementary Table S3. To retrieve missing data, or additional information, we contacted 39 authors about 35 studies. We attained raw data for 18 studies and additional information for 10 studies. Ultimately, we extracted 459 effect sizes, representing 91 experiments published in 71 articles. These studies were performed on 22 species, ranging from bees to monkeys. The main characteristics of the included studies are summarised in Figure 3, showcasing significant variation in study subjects and methodologies. Individual studies contributed between 1 and 30 effect sizes to our final data set.

Mammals were the best-represented taxonomic group (56 out of 71 studies; 330 out of 459 effect sizes), and almost all studies were performed on captive animals (65 studies; 414 effect sizes). Females were more frequently used in experiments than males or mixed-sex groups (225, 118, 116 effect sizes, respectively; for the numbers of studies see Figure 3), and adults were more commonly used than juveniles (333 and 126 effect sizes, respectively). Most often, affect manipulation was a form of stress induction compared to standard/benign conditions (benign-worse comparison: 230 effect sizes). Enrichment compared to control/benign conditions was the next most common manipulation (better-benign comparison: 135 effect sizes), and a few studies compared positive treatments (e.g. enrichment) to negative treatments (e.g. handling) (better-worse comparison: 94 effect sizes). Manipulations were usually long-term (292 effect sizes), lasting for days or weeks before affect was measured.

Between-subject designs (independent groups of animals exposed to manipulation or control/benign treatment) accounted for 302 effect sizes and within-subject designs accounted for 157 effect sizes. Go/no-go tasks dominated over active choice go/go tasks (389 and 70 effect

sizes, respectively). Spatial and visual cues were most commonly used in judgement bias tests (177 and 167 effect sizes respectively), and reward-punishment training schemes were more common than reward-null (283 and 132 effect sizes, respectively), with the remaining studies using different reward strengths (44 effect sizes). Most studies did not report whether the personnel performing measurements of animal behaviour were blinded to treatments (only 113 effect sizes came from blinded trials), or whether the measurements were automated (only 71 effect sizes came from automated trials). Finally, latency and proportion outcome measures were reported at similar levels (258 and 201 effect sizes, respectively). Only 5 studies using go/no-go tasks reported outcome measures as both latency and proportion, and these were moderately correlated (r = 0.578, t = 3.085, df = 19, p-value = 0.006), although not for the data subset using only the largest effect sizes from each experiment to remove non-independence (r = 0.443, t = 0.857, df = 3, p-value = 0.455).

An overall effect and heterogeneity among effect sizes

Overall, we found a statistically significant effect of experimental treatments on judgement bias in animals (phylogenetic multilevel meta-analysis: Hedges' g $(H_g)_{[\text{overall mean}]}$ = 0.201, 95% Confidence Interval (CI) = 0.028 to 0.374; Figure 4, Table S4). A similar model, but without controlling for phylogeny, also showed a statistically significant overall effect (multilevel meta-analysis: $H_{g[\text{overall mean}]}$ = 0.204, 95% CI = 0.087 to 0.320, Table S5). Therefore, animals in a relatively better treatment usually behaved in a more 'optimistic' way than animals in a relatively worse treatment, whereas animals in a relatively worse treatment were more 'pessimistic'. Notably, this overall effect is comparable to a small effect, as suggested by the benchmark values (0.2, 0.5 and 0.8 as small, medium and large effects; Cohen, 1969). The total heterogeneity in the whole data set was high (I^2_{total} = 76.4%; according to Higgins' benchmark 25, 50 and 75% can be interpreted as low, moderate and high heterogeneity, respectively; Higgins and Thomson, 2002). About 68.1% of the variability across studies was due to sampling

error, while phylogeny contributed little to account for this heterogeneity (2.0%), suggesting a weak phylogenetic signal (see Nakagawa & Santos, 2012).

High observed total heterogeneity in the data set warrants investigation of potential moderators of heterogeneity. We, thus, present findings of the univariate multilevel phylogenetic meta-regression models examining the effects of different moderators (see Figures 5, Figure 6, Figure S3).

Species-specific effects

A meta-regression model estimating mean effect for each included species did not show a clear pattern of differences among species (Figure 5; $R^2 = 0.070$, Table S6). Some of the species-specific point estimates were medium or large, but they were accompanied by wide confidence intervals crossing zero (no-effect) line. We note that the distribution of studies among species was not balanced, with the data set being dominated by studies on rats, cattle, and pig (15, 11 and 8 studies, respectively), while most of remaining species are each represented by a single study (Figure 5).

Sex-effects

Effects of judgement bias manipulations on males were small-to-medium and statistically different from zero ($H_{g[males]}$ = 0.365, 95% CI = 0.155 to 0.575), while effects on females were, on average, close to zero ($H_{g[females]}$ 0.104, 95% CI = -0.063 to 0.271; Figure 6a). The difference between mean effects in males and females was small ($H_{g[male\ vs.\ female\ difference]}$ = 0.261, 95% CI = -0.001 to 0.522; $R^2 = 0.024$, Table S7), indicating that affect manipulations on judgement bias measurements tend to be more pronounced in studies on males than females.

Tasks type effects

Effects of judgement bias manipulations tended to be larger in studies using active choice tasks in comparison to studies using go/no-go tasks ($H_{g[go/no\text{-}go\ vs.\ active\ choice\ difference]}$ = -0.277, 95% CI = -0.567 to 0.012; Figure 6b). On average, tasks with active choice had a medium effect size and

were statistically different from zero ($H_{g[\text{active choice}]}$ = 0.432, 95% CI = 0.151 to 0.712), while the average effect size in in go/no-go tasks was small, but still statistically different from zero ($H_{g[\text{go/no-go}]}$ = 0.154, 95% CI = 0.005 to 0.304; R^2 = 0.021, Table S8).

Cue types used during judgement bias tests

Across the five categories of cues used during judgement bias tests, only tests using auditory and tactile cues consistently revealed differences between control and affect-manipulated groups of animals ($H_{g[\text{auditory cues}]}$ = 0.393, 95% CI = 0.136 to 0.651; $H_{g[\text{tactile cues}]}$ = 0.658, 95% CI = 0.136 to 1.118; Figure 6c). These two categories of cues were only significantly different from the results from studies using visual cues, which on averaged had the weakest effect ($H_{g[\text{visual cues}]}$ = 0.067, 95% CI = -0.133 to 0.268; R^2 = 0.044, Table S9).

Reinforcement scheme during judgement bias tests

Studies using Reward-Punishment and Reward-Reward training cue reinforcement schemes usually generated small-medium statistically significant effect sizes in the predicted direction ($H_{g[\text{Reward-Punishment}]}$ = 0.216, 95% CI = 0.036 to 0.396; $H_{g[\text{Reward-Reward}]}$ = 0.488, 95% CI = 0.137 to 0.839), but not the Reward-Null reinforcement scheme (Figure 6d). Reward-Reward studies generally showed significantly larger judgement bias than those that used a Reward-Null reinforcement scheme ($H_{g[\text{Reward-Reward vs. Reward-Null}]}$ = 0.436, 95% CI = 0.045 to 0.827; R^2 = 0.030, Table S10). Studies using non-reinforced ambiguous cues (which was the vast majority of included studies) generated effect sizes in the predicted direction ($H_{g[\text{ambig. cue not reinforced}]}$ = 0.204, 95% CI = 0.026 to 0.382), although not statistically different from studies in which ambiguous cues were reinforced ($H_{g[\text{ambig. cue not reinforced vs. reinforced}]}$ = 0.080, 95% CI = -0.527 to 0.686; R^2 = 0.001), whose effect sizes were close to zero (Table S11).

Cue ambiguity level

Ambiguous cues that were halfway between the positive and negative cues, as well as cues that were closer to the negative cues, were most likely to reveal judgement bias in tested animals

($H_{g[\text{mid-point cue}]}$ = 0.250, 95% CI = 0.042 to 0.458; $H_{g[\text{near-negative cue}]}$ = 0.303, 95% CI = 0.075 to 0.530; R^2 = 0.014, Figure 6e). Ambiguous near-negative cues were also significantly different from the effects of positive training cues, with the latter on average being least likely to show judgement bias effect ($H_{g[\text{positive cues}]}$ = 0.063, 95% CI = -0.153 to 0.278, Table S12).

Other moderators in univariate models

Variation in the other considered moderators did not appear to significantly influence the magnitude of judgement bias effects. These moderators were: source of animals (captive vs. wild-caught), animal age, type of affect manipulation (stress vs. enrichment), timing of affect manipulation (short vs. long-term), whether manipulation was compared to benign or worse reference condition, type of study design (within-individual vs. between-individuals), food deprivation during judgement bias tests, measurement type of behavioural response (latency vs. proportion), automation and blinding of measurements of animal responses (Figure S3; Tables S13 – S22; R^2 = 0 to 0.010).

Multivariate (full) meta-regression models and model selection

The full meta-regression model included four moderators that were significant or close to statistical significance in univariate models (after confirming they were not co-linear with each other): sex of test animals, task type (go/no-go vs. active choice go/go), type of cue used in the test, and type of reinforcement for positive and negative training cues. In the multivariate meta-regression, none of the considered moderators was significant (Table S23). These moderators can jointly explain only about 7% of variation in the data (R^2 = 0.072). Model selection analysis indicated that type of the task and type of reinforcement used could be the most influential moderators, followed by the sex of animals (Table S24).

Publication bias

We conducted 3 kinds of publication bias analyses: 1) contour-enhanced funnel plots of residuals, 2) a variant of Egger's regression, and 3) a regression-based time-lag bias test. Visual

inspection of enhanced-contour funnel plots of residuals did not reveal skewness indicative of publication bias (Figure S4). Further, the intercept of Egger's multivariate regression, controlling for potentially important moderators from univariate models, was not significantly different from zero (t = 0.017, df = 457, p = 0.986), confirming lack of publication bias in the full data set. Finally, we found no evidence for time-lag bias, as the slope of linear regression between publication year and effect size was not significantly different from zero (Slope $_{[Year]}$ = -0.002, 95% CI = -0.121 to 0.118, p = 0.980, Table S25).

Sensitivity analyses (robustness of results)

The estimates from Bayesian models run on full data set gave qualitatively identical results to the REML models used in the main data analyses. Namely, the overall effect was small and statistically significant ($H_{g[overall\ mean]}$ = 0.206, 95% CI = 0.041 to 0.383; I^2_{total} = 76.8%; Table S26). In the Bayesian multivariate meta-regression, none of the moderators significantly influenced judgement bias test outcomes, as in the equivalent log-likelihood model.

Finally, we ran meta-analytic models on four data subsets, representing different ways of looking at the results from response curves with multiple cues: i) including only data from ambiguous cues (81 NP, 108 MID, and 80 NN effect sizes for cue locations included in this data subset), ii) including only data from mid-point ambiguous cues (108 MID effect sizes included), iii) including only data for maximum response, in absolute terms (26 P, 13 NP, 31 MID, 22 NN, and 16 N effect sizes included), iv) including only data for maximum response in the overall direction of response (19 P, 12 NP, 34 MID, 28 NN, and 15 N effect sizes included). All these data subsets tended to have larger overall effect size estimates, than in the full data set meta-analyses (Figure 4, Tables S4 and S5). Univariate and multivariate meta-regression models usually showed similar patterns to these observed in the analyses on the full dataset (Tables S6-S23).

Discussion

Our meta-analysis revealed that non-pharmacological affect manipulations generally influenced judgement bias in the predicted direction (i.e. manipulations assumed to generate a relatively positive state were likely to generate an 'optimistic' response to cues). However, effects were usually small to large (average Hedges' g of 0.2 – 0.6), and they were highly variable, with total observed heterogeneity (I^2) over 75%. The moderators that potentially influenced magnitude of effects included cue type, type of task used in judgement bias trials, reinforcement combination used for the training positive and negative cues, cue ambiguity level, and sex of tested animals. However, small R^2 values (1.4 to 4.4%) indicated that these moderators explained only small proportion of variance. We discuss these findings in detail below.

Validity and efficacy of judgement bias tests

Our main finding generally supports judgement bias tests as a valid approach to measure affect in non-human animals. This is in line with conclusions of a narrative cross-species review (Bethell, 2015) and a recent systematic review of 20 rodent studies on judgement bias (Nguyen, et al. 2020). However, the latter considered both pharmacological and non-pharmacological manipulations and only conducted a qualitative synthesis of their rodent data set. Effects of pharmacological manipulations across species were recently quantitatively synthesised by our team (Neville, et al. 2020) and our current work provides first quantification of non-pharmacological manipulations across different taxa.

Our quantitative results show that the observed behavioural effect of the affect manipulations investigated is, on average, small (Hedges' g of 0.2) and highly heterogeneous. However, we base this conclusion on the analyses of the full dataset, which included mean latency and/or proportion data from all cues used in the judgement bias tests. Thus, we likely underestimated the overall effect size, due to the inclusion of positive and negative training (unambiguous) cues, in the analysis. It is possible that the affect manipulations used in many of the included studies

were rather ”mild” – for welfare reasons, not many authors used severe stressors or pain stimuli. Some manipulations may not even have had an effect on the animal affect.

As noted earlier, there are theoretical and empirical reasons for why judgement biases may not occur at training cues, and also for why they may not occur at all ambiguous cues. When we restricted analysis to the cue with the largest absolute effect size in the direction of the overall mean effect size from each response curve – the estimated overall effect sizes were between moderate to large (Hedges' *g* of 0.6). The overall effect sizes were moderate when we used other three data subsets: (i) ambiguous cues only; (ii) middle cue only; (iii) cue with the largest absolute effect size. Yet, analyses on the full dataset are most powerful, given that they include data points representing the whole response curve (Gygax, 2014).

The high observed high data heterogeneity is congruent with the levels observed in most ecological and evolutionary meta-analyses (70 – 95%; Senior et al., 2016). High heterogeneity (> 75%) of the effect sizes in our data set indicates variability in the influences of non-pharmacological manipulations of affective state on judgement bias in animals, but is perhaps not surprising given how diverse the studies were in terms of, for example, species used (22 diverse species; Figure 5), task variants, affect manipulations, and other methodological specifics. Accordingly, the lack of phylogenetic effects in our data set is consistent with the observation that meta-analyses on phylogenetically diverse sets of species are unlikely to show a strong phylogenetic signal (Chamberlain et al., 2012).

Key moderators of judgement bias tests

We have also revealed five important moderators of responses in the judgement bias task. Four are related to methodology and one is a biological factor. First, active choice go/go tasks tended to yield larger effects than go/no-go tasks. It is possible that the former are more cognitively challenging given that the response needs to be deployed to different stimuli. Such a potential cognitive load might render go/go tasks less susceptible to habitual responding and thereby more sensitive to affect manipulations. Furthermore, go/no-go tasks are likely to be vulnerable

to the influence of Pavlovian action predispositions (e.g. go-for-reward; no-go to avoid punishment; Guitart-Masip et al., 2014; Jones et al., 2017), that could inadvertently bias responding (Mendl & Paul, 2020) and obscure affect manipulation effects. Additionally, subjects may sometimes perform no-go responses for reasons unrelated to affect manipulations (e.g. failing to detect or attend to a cue; Bethell, 2015; Jones et al., 2018), making these tests less dependable. Still, we observed that go/no-go tasks are more commonly used in judgement bias studies (in 57 vs. 14 studies; Figure 3), probably because they are easier and quicker to train.

Second, Reward-Reward tasks usually generated larger effect sizes than Reward-Null tasks. Part of the reason for this may be that Reward-Reward tasks usually involve a go/go active choice response and this itself predisposes stronger effects, as just discussed. The most frequently used Reward-Punishment tasks had a largest observed average effect size. It is possible that Reward-Punishment design, providing a more affectively-laden task (i.e. decision outcomes can range from a desired reward to an aversive punisher), is more sensitive to manipulations of affective state (see Mendl et al. 2009).

Third, the use of auditory and tactile cues tended to reveal the largest effects compared to when spatial, visual and olfactory cues were employed. There may be a number of reasons for this, some of which may be linked to differences in species biology (Bethell, 2015). For example, whilst people are strongly visually focused when information gathering, many other animal species are not, and may not readily exhibit human-like processing of visual cues. Conversely, olfactory sensitivity in humans is poor, relative to many other species, and this may impair the ability of researchers to design or use meaningful cues in this sensory dimension. It is also possible that cue modality and presentation method can influence the uncertainty of information provided by 'ambiguous' cues. For example, there may be greater uncertainty about the information provided by a single tone intermediate between two training tones, than by a spatial location situated between two training locations. Such differences in uncertainty may have knock-on effects on animal's decisions.

Fourth, cue ambiguity level (P, NP, MID, NN, N) was important. We found predicted judgement bias only at ambiguous cues in the full dataset analysis, and not at positive or negative training cues, on average. Still, some individual studies in the dataset yielded large effects at positive or negative training cues (e.g. Deakin, 2018; Horváth et al., 2016; Zidar et al., 2018). In line with this, Neville et al. (2020) noted that pharmacological manipulations of affect altered judgement bias principally at ambiguous cues, but also at the negative training cue. Large effects at non-ambiguous cues could occur in at least two ways. First, if affect manipulations altered valuation of decision outcomes (e.g. by decreasing food valuation and hence generating a weaker response to the positive cue), the manipulations could change propensities to perform specific responses (e.g. go vs. no-go) and interfere with memory of training cue-outcome associations. Second, large effects at non-ambiguous cues might occur if training was brief or ineffective such that there was considerable ambiguity about the training cue-outcome association during testing (see Mendl et al., 2009; Bateson et al. 2011; Bethell, 2015; Mendl and Paul, 2020).

Finally, in all analyses, larger predicted effect sizes tended to be reported for male subjects than for females or mixed sex groups. This pattern could be due to existence of sex differences in neurobiology of learning and memory (Jonasson et al., 2005) or sex differences in stress effects on memory, with different patterns for acute and prolonged stress (Andreano and Cahill, 2009). Effects of enrichment may also be sex-specific (Lin et al., 2011; ter Horst et al., 2012).

Potential limitations and recommendations

The results of our meta-analysis come with six caveats, which we list here alongside recommendations for future studies of judgement bias. First, captive and domesticated mammals dominate the dataset making our conclusions particularly relevant to research on welfare of such animals. Conversely, the analyses are less informative for wild animals, vertebrates other than mammals, and invertebrates. Indeed, Bethell's narrative review (2015) highlighted biased taxonomic representation in empirical evidence. Thus, future work in this area could aim to increase the representation of non-domesticated species, such as these kept in

zoos and for research (where animal welfare is of concern; Baumans, 2005; Bethell, 2015; Wolfensohn et al., 2018) and invertebrates (where welfare is an emerging issue; Drinkwater et al., 2019).

Second, we had limited statistical power to detect clear differences between the levels of number of the tested moderators. Also, the small sample sizes at some levels of the considered moderators might have introduced some spurious findings. For example, relatively few studies used tactile or olfactory cues (e.g. in Barker et al., 2017; Novak et al., 2016), and very few used reinforced ambiguous cues during tests (e.g. in Bailoo et al., 2018; Keen et al., 2014). To address this limitation, future studies of commonly used laboratory and domesticated species should systematically investigate the role of different cue types. Researchers should also attempt to make cue types relevant for a given species, and vary the perceptual closeness of training cues and hence the difficulty of the task and uncertainty of ambiguous cues.

Third, for some moderators, especially these related to study quality, poor reporting might have obscured statistical relationships. Very few of the included studies explicitly stated that they used automation or blinding, and we had to assume that the remaining studies did not use these. Thus, automation of measurements could be used more often, and/or their use should be clearly reported. Notably, Nguyen et al. (2020) in their systematic review of 20 rodent studies highlighted limited information on the details of experimental procedures and analyses in 65% of assessed studies, undermining confidence in the findings. Nevertheless, we found no statistical evidence for publication bias in our meta-analytic data set. The lack of publication bias is potentially due to our full data set containing data points across the whole response curve, which are usually a mixture of small and large positive effects (in the expected direction) and even some negative ones (not in the expected direction). Also related to reporting, mixed-sex groups of animals comprised almost one-third of the data in our meta-analysis, potentially obscuring sex-specific effects. Providing sex-disaggregated data in research is absolutely

essential for improving our understanding of animal behaviour and cognition (Shansky and Woolley, 2016; Palanza and Parmigiani, 2017).

Fourth, we were also not able to include strength of manipulation in our analyses (there is no common scale for the diverse types of manipulations included in our data set). To overcome this problem, in future studies it would be valuable to test and synthesize relationships between measures of cognitive bias and different biomarkers of stress, such as cortisol, adrenaline, alpha-amylase, testosterone, leucocyte profiles (Keay et al., 2006; Davis et al., 2008).

Fifth, we also noted some outliers in the dataset, which usually came from studies with severe manipulations and/or small sample sizes. We, however, conducted extensive sensitivity analyses to test robustness of our conclusions, with the results generally conforming to our predictions and being robust across different statistical approaches. Further, in individual empirical studies comparing two means, to achieve power of 0.8 at alpha of 0.05, it is necessary to have sample sizes of at least 50 animals per group for detecting moderate effect sizes (Hedges' g = 0.4). Best-case scenario, when effect is large (Hedges' g = 0.8), would require only 13 animals per group to achieve the same power. Conducting power analyses to determine suitable sample sizes for planned experiments can help reducing animal use and also prevent wasting animals on underpowered studies.

Finally, the largest responses often do not appear at the most intermediate/ambiguous cue. Because of this, we suggest that multiple ambiguous (probe) cues (at least 3) are needed for robust and comprehensive judgement bias tests although 25% of response curves in our data set included only one ambiguous cue.

Conclusions

In summary, judgement bias tests are a valid method of measuring animal affective state. However, high heterogeneity among studies, which can be only partially explained by simple

influences of considered moderators, warrants care in designing and interpreting judgement bias manipulations and tests. We call for better reporting of experimental designs, especially blinding and automation, disaggregation of data by sex of subjects, and other experimental details that might influence study results. Also, there is a need for more empirical studies that compare different experimental designs and setups, including using different types of tasks, cues, and cue ambiguity levels.

Acknowledgments

We are thankful to all authors that have provided us with additional information or data, as indicated in Table S2. M.L. and S.N. were supported by the Australian Research Council Discovery Project (DP200100367). E.S. and J.Z. were funded by Carl Trygger's Foundation and the Swedish research council Formas, respectively, awarded to H.L.. M.M., E.S.P. and V.N. thank the UK Biotechnology and Biological Sciences Research Council (BBSRC grants BB/P019218/1, BB/T002654/1 and BBSRC SWBio DTP grant BB/M009122/1), and the UK National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs grant NC/K00008X/1) for supporting their work in this area. The authors declare no conflicts of interests.

Data and Code Availability

All data and code are available from the following online OSF repository: https://osf.io/anfhm/

References

Andreano, J.M., Cahill, L., 2009. Sex influences on the neurobiology of learning and memory. Learn. Mem. 16, 248–266. https://doi.org/10.1101/lm.918309

Ash, H., Buchanan-Smith, H.M., 2016. The long-term impact of infant rearing background on the affective state of adult common marmosets (Callithrix jacchus). Appl. Anim. Behav. Sci. 174, 128–136. https://doi.org /10.1016/j.applanim.2015.10.009

Baciadonna, L., McElligott, A.G., 2015. The use of judgement bias to assess welfare in farm livestock. Anim. Welf. 24, 81–91. https://doi.org/10.7120/09627286.24.1.081

Bailoo, J.D., Murphy, E., Boada-Saña, M., Varholick, J.A., Hintze, S., Baussière, C., Hahn, K.C., Göpfert, C., Palme, R., Voelkl, B., Würbel, H., 2018. Effects of cage enrichment on behavior, welfare and outcome variability in female mice. Front. Behav. Neurosci. 12, 232. https://doi.org/10.3389/fnbeh.2018.00232.

Barker, T.H., Bobrovskaya, L., Howarth, G.S., Whittaker, A.L., 2017. Female rats display fewer optimistic responses in a judgment bias test in the absence of a physiological stress response. Physiol. Behav. 173, 124–131. https://doi.org/10.1016/j.physbeh.2017.02.006

Barton, K., 2009. MuMIn: Multi-model inference. R Package Version 0.12.2/r18. http://R-Forge.R-project.org/projects/mumin/

Bateson, M., 2016. Optimistic and pessimistic biases: a primer for behavioural ecologists. Curr. Opin. Behav. Sci. 12, 115–121. https://doi.org/10.1016/j.cobeha.2016.09.013

Bateson, M., Desire, S., Gartside, S. E., Wright, G. A., 2011. Agitated honeybees exhibit pessimistic cognitive biases. Curr. Biol. 21, 1070–1073. doi: 10.1016/j.cub.2011.05.017

Baumans, V., 2005. Science-based assessment of animal welfare: laboratory animals. Rev. Sci. Tech. 24, 503-513. http://dx.doi.org/10.20506/rst.24.2.1585

Bethell, E.J., 2015. A "How-To" Guide for Designing Judgment Bias Studies to Assess Captive Animal Welfare. J. Appl. Anim. Welf. Sci. 18, S18–S42. https://doi.org/ 10.1080/10888705.2015.1075833.

Bethell, E.J., Holmes, A., MacLarnon, A., Semple, S., 2016. Emotion Evaluation and Response Slowing in a Non-Human Primate: New Directions for Cognitive Bias Measures of Animal Emotion? Behav. Sci. (Basel). 6, 2. https://doi.org/10.3390/bs6010002

Bethell, E.J., Koyama, N.F., 2015. Happy hamsters? Enrichment induces positive judgement bias for mildly (but not truly) ambiguous cues to reward and punishment in Mesocricetus auratus. R. Soc. Open Sci. 2, 140399. https://doi.org/10.1098/rsos.140399

Blanchette, I., Richards, A., 2010. The influence of affect on higher level cognition: A review of research on interpretation, judgement, decision making and reasoning. Cogn. Emot. https://doi.org/10.1080/02699930903132496

Brilot, B.O., Asher, L., Bateson, M., 2010. Stereotyping starlings are more "pessimistic." Anim. Cogn. 13, 721–731. https://doi.org/10.1007/s10071-010-0323-z

Burman, O.H.P., Mendl, M.T., 2018. A novel task to assess mood congruent memory bias in non-human animals. J. Neurosci. Methods 308, 269–275.

https://doi.org/10.1016/j.jneumeth.2018.07.003

Burman, O.H.P., Parker, R.M.A., Paul, E.S., Mendl, M.T., 2009. Anxiety-induced cognitive bias in non-human animals. Physiol. Behav. 98, 345–350. https://doi.org/10.1016/j.physbeh.2009.06.012

Chamberlain, S.A., Hovick, S.M., Dibble, C.J., Rasmussen, N.L., Van Allen, B.G., Maitner, B.S., Ahern, J.R., Bell‐Dereske, L.P., Roy, C.L., Meza‐Lopez, M., Carrillo, J., Siemann, E., Lajeunesse, M.J. and Whitney, K.D., 2012. Does phylogeny matter? Assessing the impact of phylogenetic information in ecological meta‐analysis. Ecol. Lett. 15, 627-636. doi:10.1111/j.1461-0248.2012.01776.x

Crump, A., Arnott, G., Bethell, E.J., 2018. Affect-Driven Attention Biases as Animal Welfare Indicators: Review and Methods. ANIMALS 8. https://doi.org/10.3390/ani8080136

da Cunha Nogueira, S.S., Fernandes, I.K., Oliveira Costa, T.S., Gama Nogueira-Filho, S.L., Mendl, M., 2015. Does Trapping Influence Decision-Making under Ambiguity in White-Lipped Peccary (Tayassu pecari)? PLOS One 10. https://doi.org/10.1371/journal.pone.0127868

Cohen, J. (1969). Statistical power analysis for the behavioral sciences (1st ed.). New York, NY: Academic Press.

Darwin, C., 1872. The expression of the emotions in man and animals, 3rd ed., The expression of the emotions in man and animals, 3rd ed. Oxford University Press, New York, NY, US.

Davis, A.K., Maney, D.L., Maerz, J.C., 2008. The use of leukocyte profiles to measure stress in vertebrates: a review for ecologists. Funct. Ecol. 22: 760-772. https://doi.org/10.1111/j.1365-2435.2008.01467.x

Destrez, A., Deiss, V., Lévy, F., Calandreau, L., Lee, C., Chaillou-Sagon, E., Boissy, A., 2013. Chronic stress induces pessimistic-like judgment and learning deficits in sheep. Appl. Anim. Behav. Sci. 148, 28–36. https://doi.org/10.1016/j.applanim.2013.07.016

Douglas, C., Bateson, M., Walsh, C., Bédué, A., Edwards, S.A., 2012. Environmental enrichment induces optimistic cognitive biases in pigs. Appl. Anim. Behav. Sci. 139, 65–73. https://doi.org/10.1016/j.applanim.2012.02.018

Doyle, R.E., Lee, C., Deiss, V., Fisher, A.D., Hinch, G.N., Boissy, A., 2011. Measuring judgement bias and emotional reactivity in sheep following long-term exposure to unpredictable and aversive events. Physiol. Behav. 102, 503–510. https://doi.org/10.1016/j.physbeh.2011.01.001

Drinkwater, E., Robinson, E.J.H., Hart, A.G., 2019. Keeping invertebrate research ethical in a landscape of shifting public opinion. Methods Ecol Evol. 10, 1265–1273.

https://doi.org/10.1111/2041-210X.13208

Egger, M., Smith, G.D., Schneider, M., Minder, C., 1997. Bias in meta-analysis detected by a simple , graphical test measures of funnel plot asymmetry. BMJ 315, 629–634. https://doi.org/10.1136/bmj.315.7109.629

Girard, B., Bellone, C., 2020. Revealing animal emotions. Science 368, 33–34. doi: 10.1126/science.abb2796

Guitart-Masip M., Duzel, E., Dolan, R., Dayan, P., 2014. Action versus valence in decision making. Trends Cogn. Sci. 18,194-202. https://doi.org/10.1016/j.tics.2014.01.003

Gygax, L., 2014. The A to Z of statistics for testing cognitive judgement bias. Anim. Behav. 95, 59–69. https://doi.org/10.1016/j.anbehav.2014.06.013

Hadfield, J.D., 2010. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. J. Stat. Softw. 33, 1–22. https://doi.org/10.18637/jss.v033.i02

Hales, C.A., Robinson, E.J., Houghton, C.J., 2016. Diffusion modelling reveals the decision making processes underlying negative judgement bias in rats. PLOS One 11. https://doi.org/10.1371/journal.pone.0152592

Hales, C.A., Stuart, S.A., Anderson, M.H., Robinson, E.S., 2014. Modelling cognitive affective biases in major depressive disorder using rodents. Br. J. Pharmacol. 171, 4524–4538. https://doi.org/10.1111/bph.12603

Harding, E.J., Paul, E.S., Mendl, M., 2004. Cognitive bias and affective state. Nature 427, 312.

Hedges, L. V., Olkin, I., 1985. Statistical Methods for Meta-Analysis, Statistical Methods for Meta-Analysis. New York: Academic Press. https://doi.org/10.1016/c2009-0-03396-0

Jennions, M.D., Møller, A.P., 2002. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proceedings. Biol. Sci. 269, 43–8. https://doi.org/10.1098/rspb.2001.1832

Jonasson, Z., 2005. Meta-analysis of sex differences in rodent models of learning and memory: A review of behavioral and biological data. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2004.10.006

Jones, S., Neville, V., Higgs, L., Paul, E.S., Dayan, P., Robinson, E.S.J., Mendl, M,, 2018. Assessing animal affect: an automated and self-initiated judgement bias task based on natural investigative behaviour. Scientific Reports 8. doi: 10.1038/s41598-018-30571-x

Jones, S., Paul, E.S., Dayan, P., Robinson, E.S.J., Mendl, M., 2017. Pavlovian influences on learning differ between rats and mice in a counter-balanced Go/NoGo judgement bias task. Behavioural Brain Research 331:214-224. doi: 10.1016/j.bbr.2017.05.044

Keay, J.M., Singh, J., Gaunt, M.C., Kaur, T., 2006. Fecal glucocorticoids and their metabolites as indicators of stress in various mammalian species: a literature review. J. Zoo Wildlife Med., 37, 234-244. https://doi.org/10.1638/05-050.1

Keen, H.A., Nelson, O.L., Robbins, C.T., Evans, M., Shepherdson, D.J., Newberry, R.C., 2014. Validation of a novel cognitive bias task based on difference in quantity of reinforcement for assessing environmental enrichment. Anim. Cogn. 17, 529–541. https://doi.org/10.1007/s10071-013-0684-1

Lin, E.J., Choi, E., Liu, X., Martin, A., & During, M.J. (2011). Environmental enrichment exerts sex-specific effects on emotionality in C57BL/6J mice. Behav. Brain Res. 216, 349–357. https://doi.org/10.1016/j.bbr.2010.08.019

Mendl, M., Burman, O.H.P., Parker, R.M.A., Paul, E.S., 2009. Cognitive bias as an indicator of animal emotion and welfare: Emerging evidence and underlying mechanisms. Appl. Anim. Behav. Sci. 118, 161–181. https://doi.org/10.1016/j.applanim.2009.02.023

Mendl, M., Burman, O.H.P., Paul, E.S., 2010. An integrative and functional framework for the study of animal emotion and mood. Proc. R. Soc. B Biol. Sci. 277, 2895–2904. https://doi.org/10.1098/rspb.2010.0303

Mendl, M., Paul, E.S. 2020. Animal affect and decision-making. Neurosci. Biobehav. Rev. 112, 144-163. https://doi.org/10.1016/j.neubiorev.2020.01.025

Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Altman, D., Antes, G., Atkins, D., Barbour, V., Barrowman, N., Berlin, J.A., Clark, J., Clarke, M., Cook, D., D'Amico, R., Deeks, J.J., Devereaux, P.J., Dickersin, K., Egger, M., Ernst, E., Gøtzsche, P.C., Grimshaw, J., Guyatt, G., Higgins, J., Ioannidis, J.P.A., Kleijnen, J., Lang, T., Magrini, N., McNamee, D., Moja, L., Mulrow, C., Napoli, M., Oxman, A., Pham, B., Rennie, D., Sampson, M., Schulz, K.F., Shekelle, P.G., Tovey, D., Tugwell, P., 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Med. https://doi.org/10.1371/journal.pmed.1000097

Murphy, E., Nordquist, R.E., van der Staay, F.J., 2013. Responses of conventional pigs and Gottingen miniature pigs in an active choice judgement bias task. Appl. Anim. Behav. Sci. 148, 64–76. https://doi.org/10.1016/j.applanim.2013.07.011

Nakagawa, S., Johnson, P.C.D., Schielzeth, H., 2017. The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. J. R. Soc. Interface 14. https://doi.org/10.1098/rsif.2017.0213

Nakagawa, S., Lagisz, M., O'Dea, R.E., Rutkowska, J., Yang, Y., Noble, D., Senior, A.M., 2020. The Orchard Plot: Cultivating Forest Plots for Use in Ecology, Evolution and Beyond. https://doi.org/10.32942/OSF.IO/EPQA7

Nakagawa, S., Santos, E.S.A., 2012. Methodological issues and advances in biological meta-analysis. Evol. Ecol. https://doi.org/10.1007/s10682-012-9555-5

Nakagawa, S., Schielzeth, H., 2013. A general and simple method for obtaining R2 from generalized linear mixed-effects models. Meth. Ecolol. Evol. 4, 133–142 https://doi.org/10.1111/j.2041-210x.2012.00261.x

Neville, V., Nakagawa, S., Zidar, J., Paul, E.S., Lagisz, M., Bateson, M., Løvlie, H., Mendl, M., 2020. Pharmacological manipulations of judgement bias: A systematic review and meta-analysis. Neurosci. Biobehav. Rev. 108, 269–286. https://doi.org/10.1016/j.neubiorev.2019.11.008

Nguyen, H.A.T., Guo, C., Homberg, J.R., 2020. Cognitive bias under adverse and rewarding conditions: a systematic review of rodent studies. Front. Behav. Neurosci. 14:14. doi: 10.3389/fnbeh.2020.00014

Novak, J., Stojanovski, K., Melotti, L., Reichlin, T.S., Palme, R., Würbel, H., 2016. Effects of stereotypic behaviour and chronic mild stress on judgement bias in laboratory mice. Appl. Anim. Behav. Sci. 174, 162–172. https://doi.org/10.1016/j.applanim.2015.10.004

Oliveira, F.R.M., Nogueira, S.L.G., Sousa, M.B.C., Dias, C.T.S., Mendl, M., Nogueira, S.S.C., 2016. Measurement of cognitive bias and cortisol levels to evaluate the effects of space restriction on captive collared peccary (Mammalia, Tayassuidae). Appl. Anim. Behav. Sci. 181, 76–82. https://doi.org/10.1016/j.applanim.2016.05.021

Palanza, P., Parmigiani, S., 2017. How does sex matter? Behavior, stress and animal models of neurobehavioral disorders. 76A: 134-143. Neurosci. Biobehav. Rev. https://doi.org/10.1016/j.neubiorev.2017.01.037

Paul, E.S., Harding, E.J., Mendl, M., 2005. Measuring emotional processes in animals: The utility of a cognitive approach. Neurosci. Biobehav. Rev. 29, 469–491. https://doi.org/10.1016/j.neubiorev.2005.01.002

Paul, E.S., Sher, S., Tamietto, M., Winkielman, P., Mendl, M.T., 2020. Towards a comparative science of emotion: Affect and consciousness in humans and animals. Neurosci. Biobehav. Rev. https://doi.org/10.1016/J.NEUBIOREV.2019.11.014

R Development Core Team (2018) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna.

Roelofs, S., Boleij, H., Nordquist, R.E., van der Staay, F.J., 2016. Making decisions under ambiguity: Judgment bias tasks for assessing emotional state in animals. Front. Behav. Neurosci. 10.

Rolls, E.T., 2009. Emotion Explained, Emotion Explained. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198570035.001.0001

Rosenthal, R., 1979. The file drawer problem and tolerance for null results. Psychol. Bull. 86, 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Rygula, R., Papciak, J., Popik, P., 2013. Trait pessimism predicts vulnerability to stress-induced anhedonia in rats. Neuropsychopharmacology 38, 2188–2196. https://doi.org/10.1038/npp.2013.116

Scherer, K.R., 2005. What are emotions? And how can they be measured? Soc. Sci. Inf. 44, 695–729. https://doi.org/10.1177/0539018405058216

Senior, A.M., Grueber, C.E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E.S.A., Nakagawa, S., 2016. Heterogeneity in ecological and evolutionary meta-analyses: Its magnitude and implications. Ecology 97. https://doi.org/10.1002/ecy.1591

Shansky, R.M., Woolley, C.S., 2016. Considering sex as a biological variable will be valuable for neuroscience research. J. Neurosci. 36, 11817–11822. https://doi.org/10.1523/JNEUROSCI.1390-16.2016

ter Horst, J.P., de Kloet, E.R., Schächinger, H., & Oitzl, M S. (2012). Relevance of stress and female sex hormones for emotion and cognition. Cell. Mol. Neurobiol. 32, 725–735. https://doi.org/10.1007/s10571-011-9774-2

Viechtbauer, W., 2010. Conducting Meta-Analyses in R with the metafor Package. J. Stat. Softw. https://doi.org/10.1103/PhysRevB.91.121108

Wallace, B.C., Small, K., Brodley, C.E., Lau, J., Trikalinos, T.A., 2012. Deploying an interactive machine learning system in an Evidence-based Practice Center: Abstrackr, in: IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. pp. 819–823. https://doi.org/10.1145/2110363.2110464

Wheeler, R.R., Swan, M.P., Hickman, D.L., 2015. Effect of multilevel laboratory rat caging system on the well-being of the singly-housed sprague dawley rat. Lab. Anim. 49, 10–19. https://doi.org/10.1177/0023677214547404

Wolfensohn, S., Shotton, J., Bowley, H., Davies, S., Thompson, S., Justice, W.S.M., 2018. Assessment of Welfare in Zoo Animals: Towards Optimum Quality of Life. Animals 8, 110. https://doi.org/ 10.3390/ani8070110

Zidar, J., Campderrich, I., Jansson, E., Wichman, A., Winberg, S., Keeling, L., Løvlie, H., 2018. Environmental complexity buffers against stress-induced negative judgement bias in female chickens. Sci. Rep. 8, 5404. https://doi.org/10.1038/s41598-018-23545-6

Figure captions

Figure 1

Conceptual diagram presenting the main elements of a typical judgement bias study. a) The basic task is trained using either a go/no-go, or active choice (go/go) design. b) Manipulations of affective state usually, but not always, occur after training of the task and may be acute or longer-term. c) Tests involve the standard training protocol plus the addition of occasionally presented ambiguous cues whose properties are usually intermediate between the trained positive and negative cues (NP = near positive cue, MID = intermediate between positive and negative cue, NN = near negative cue). d) 'Optimistic' and 'pessimistic' responding to the cues is inferred from the proportion of positive responses and/or the latency to make positive responses, which are inversely related.

Figure 2

PRISMA flow diagram. Articles identified and number of articles included and excluded during each screening stage.

Figure 3

Main characteristics of the included studies. Blue bars represent numbers of studies represented in each level of categorical variables. Between one and 30 effect sizes were extracted per study and the distribution of effect sizes generally follows the pattern of the presented data aggregated to the study level (e.g. some studies reported data for only one sex, others reported data for both sexes together, and 3 studies that included both sexes reported data for females and males separately, shown here as 'female and male sep.'). Numbers do not add up to 71 for some of the variables due to multiple experiments being present within some studies, or complex experimental designs being used.

Figure 4

Forest-like (orchard) plots showing effect size (Hedges' g) estimates from meta-analyses on: a) whole data set (all cues reported for judgement bias tests), and b-e) four subsets of this data set, representing different ways of interpreting the judgement bias test results. Positive effect sizes indicate a positive effect of affect manipulation treatments on judgement bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulations working in

the expected direction. The effects are statistically significant when the thick horizontal error bars (95% confidence intervals) do not cross zero. Thin horizontal whiskers indicate prediction intervals. *k* is number of effect sizes. Dots represent individual effect sizes scaled proportionally to their precision.

Figure 5

Forest plot showing mean effect size (Hedges' *g*) estimates from meta-regression analysis using species identity as a moderator. Positive effect sizes indicate a positive effect of affect manipulation treatments on judgment bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulations working in the expected direction. The effects are statistically significant when the horizontal error bars (95% confidence intervals) do not cross zero. *k* is number of effect sizes, *K* is number of studies.

Figure 6

Forest plots showing effect size (Hedges' *g*) estimates from the univariate meta-regression analyses (one moderator at a time) with potentially influential moderators. Effect sizes with positive values indicate a positive effect of affect manipulations on judgement bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulation treatment working in the expected direction. The mean effects (black unfilled circles) for each group of individual effect sizes (grey filled circles) are statistically different from zero when their horizontal error bars (95% confidence intervals) do not cross zero. Thin horizontal whiskers indicate prediction intervals. *k* is number of effect sizes. Dots represent individual effect sizes scaled proportionally to their precision.

Figure1

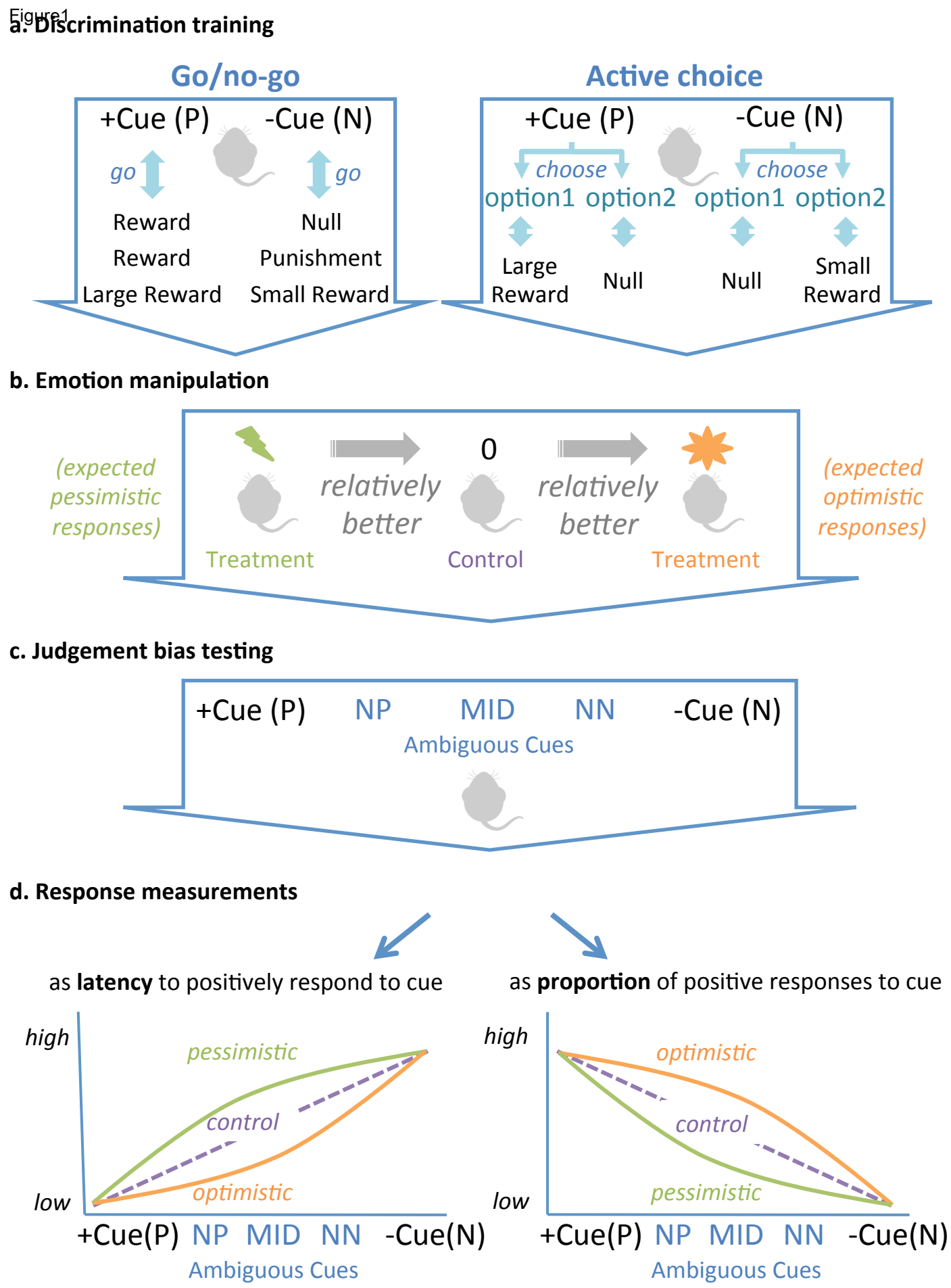

a. Discrimination training
Go/no-go
+Cue (P)
-Cue (N)
go
go
Reward
Reward
Large Reward
Null
Punishment
Small Reward
Active choice
+Cue (P)
-Cue (N)
choose
choose
option1
option2
option1
option2
Large Reward
Null
Null
Small Reward
b. Emotion manipulation
(expected pessimistic responses)
relatively better
0
relatively better
(expected optimistic responses)
Treatment
Control
Treatment
c. Judgement bias testing
+Cue (P)
NP
MID
NN
-Cue (N)
Ambiguous Cues
d. Response measurements
as latency to positively respond to cue
as proportion of positive responses to cue
high
low
pessimistic
control
optimistic
+Cue(P)
NP
MID
NN
-Cue(N)
Ambiguous Cues
high
low
optimistic
control
pessimistic
+Cue(P)
NP
MID
NN
-Cue(N)
Ambiguous Cues

Figure2

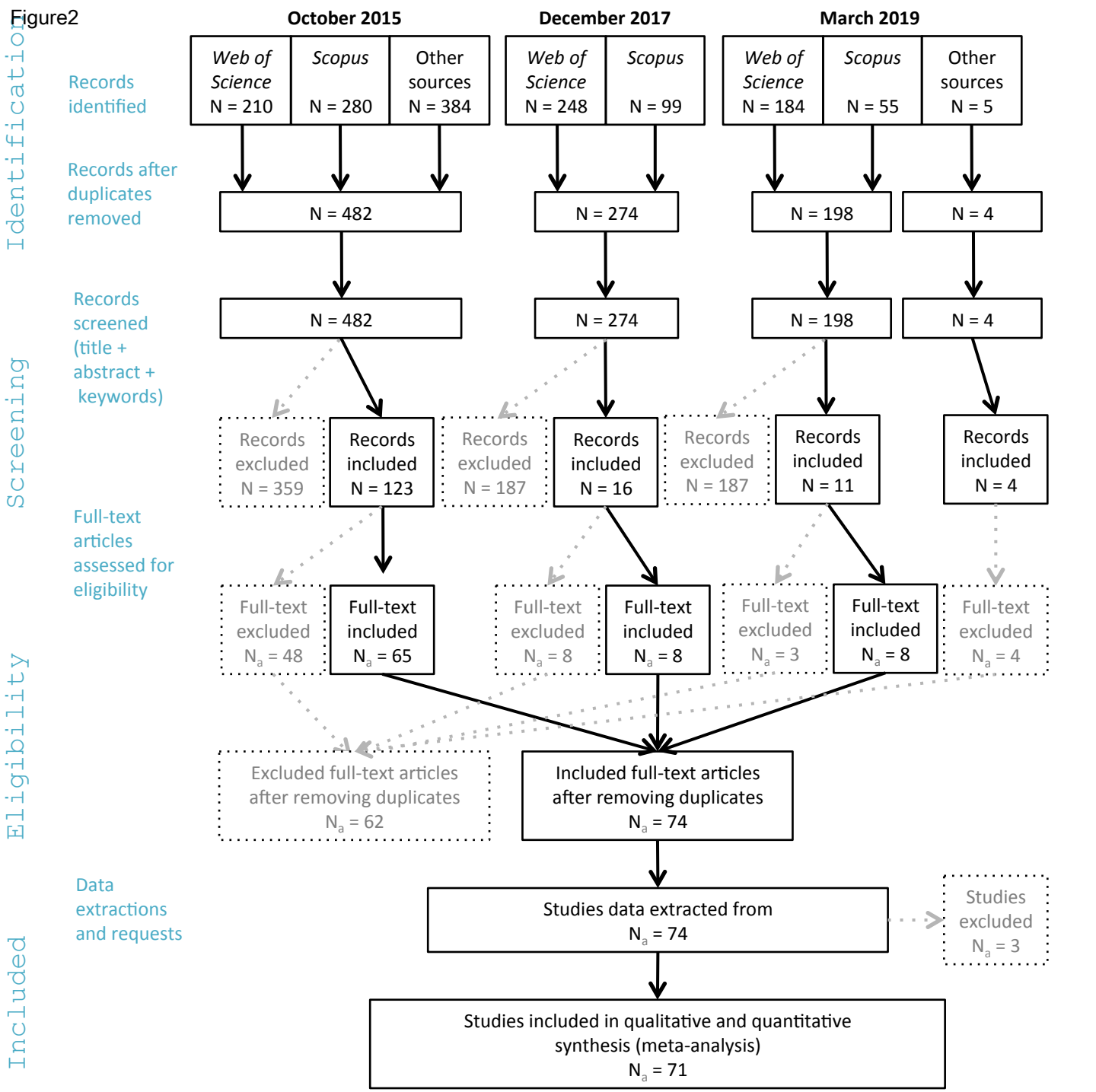

October 2015
December 2017
March 2019
Identification
Screening
Eligibility
Included
Records identified
Web of Science N = 210
Scopus N = 280
Other sources N = 384
Web of Science N = 248
Scopus N = 99
Web of Science N = 184
Scopus N = 55
Other sources N = 5
Records after duplicates removed
N = 482
N = 274
N = 198
N = 4
Records screened (title + abstract + keywords)
N = 482
N = 274
N = 198
N = 4
Records excluded N = 359
Records included N = 123
Records excluded N = 187
Records included N = 16
Records excluded N = 187
Records included N = 11
Records included N = 4
Full-text articles assessed for eligibility
Full-text excluded Na = 48
Full-text included Na = 65
Full-text excluded Na = 8
Full-text included Na = 8
Full-text excluded Na = 3
Full-text included Na = 8
Full-text excluded Na = 4
Excluded full-text articles after removing duplicates Na = 62
Included full-text articles after removing duplicates Na = 74
Data extractions and requests
Studies data extracted from Na = 74
Studies excluded Na = 3
Studies included in qualitative and quantitative synthesis (meta-analysis) Na = 71

Figure3

71 included studies

Taxon of animals
- 56 mammal
- 11 bird
- 4 insect

Origin of animals
- 65 captive
- 6 wild-caught

Sex of animals
- 27 only female
- 29 only male
- 22 mixed-sex groups
- 3 female and male sep.

Age of animals
- 53 adults / no info
- 18 juveniles

Treatment type
- 49 stress
- 22 enrichment

Comparison type
- 39 Benign-Worse
- 18 Better-Benign
- 18 Better-Worse

Treatment timing
- 44 long-term
- 29 before/during

Study design
- 49 between
- 12 within (crossover)
- 12 within (before-after)

Task type
- 57 go/no-go
- 14 active choice

Cue type
- 33 spatial
- 18 visual
- 12 auditory
- 5 tactile
- 3 olfactory

Training reinforcement
- 43 Reward-Punishment
- 18 Reward-Null
- 10 Reward-Reward(small)

Ambiguous cues reinforced
- 68 no / no info
- 3 yes

Food deprived
- 52 no / no info
- 19 yes

Measure type
- 36 latency
- 40 proportion

Measurement automation
- 63 no / no info
- 8 yes

Personnel blinding
- 54 no / no info
- 17 yes

Figure4

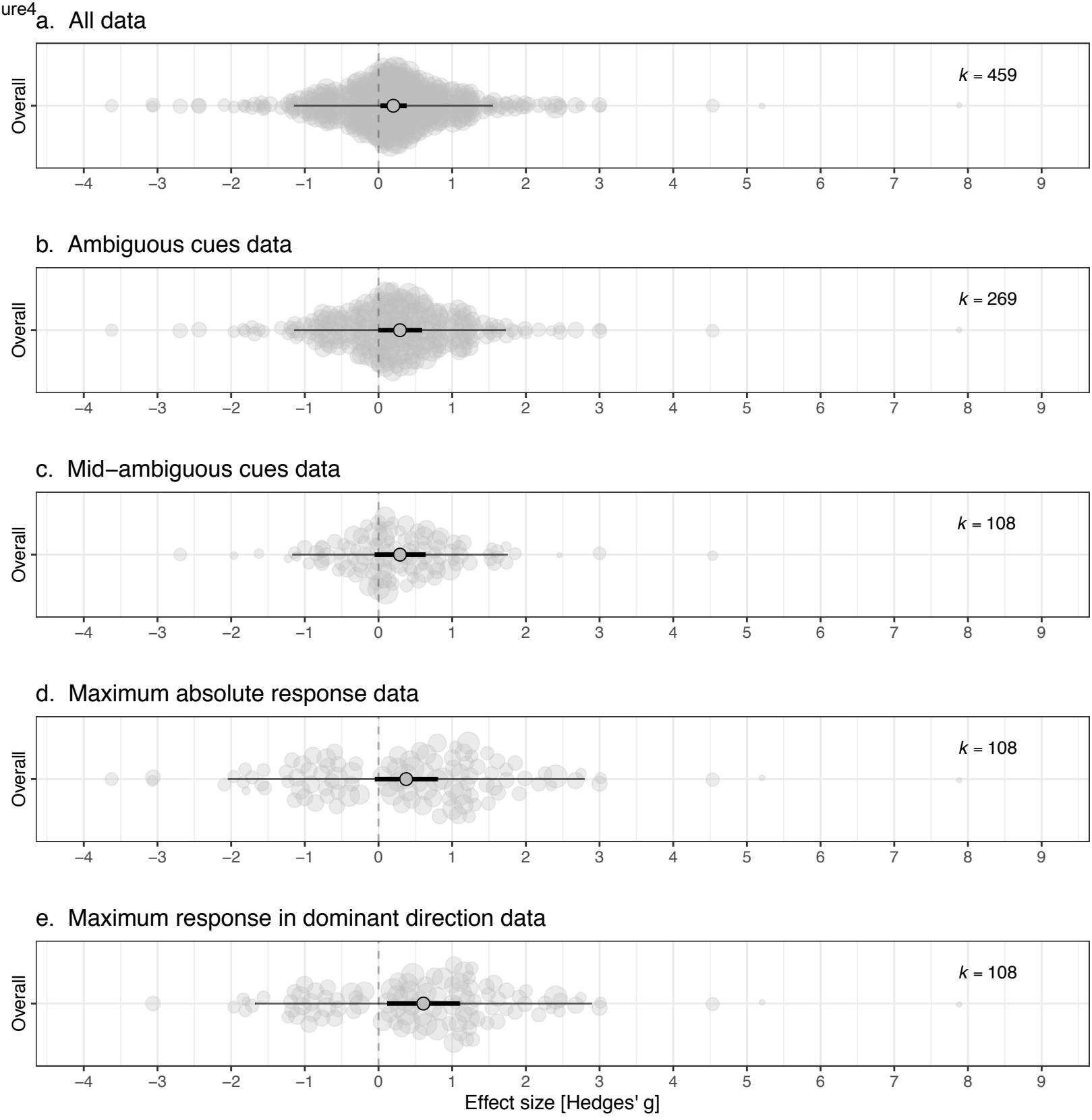

a. All data
Overall
k = 459
b. Ambiguous cues data
Overall
k = 269
c. Mid−ambiguous cues data
Overall
k = 108
d. Maximum absolute response data
Overall
k = 108
e. Maximum response in dominant direction data
Overall
k = 108
−4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9
Effect size [Hedges' g]

Figure5

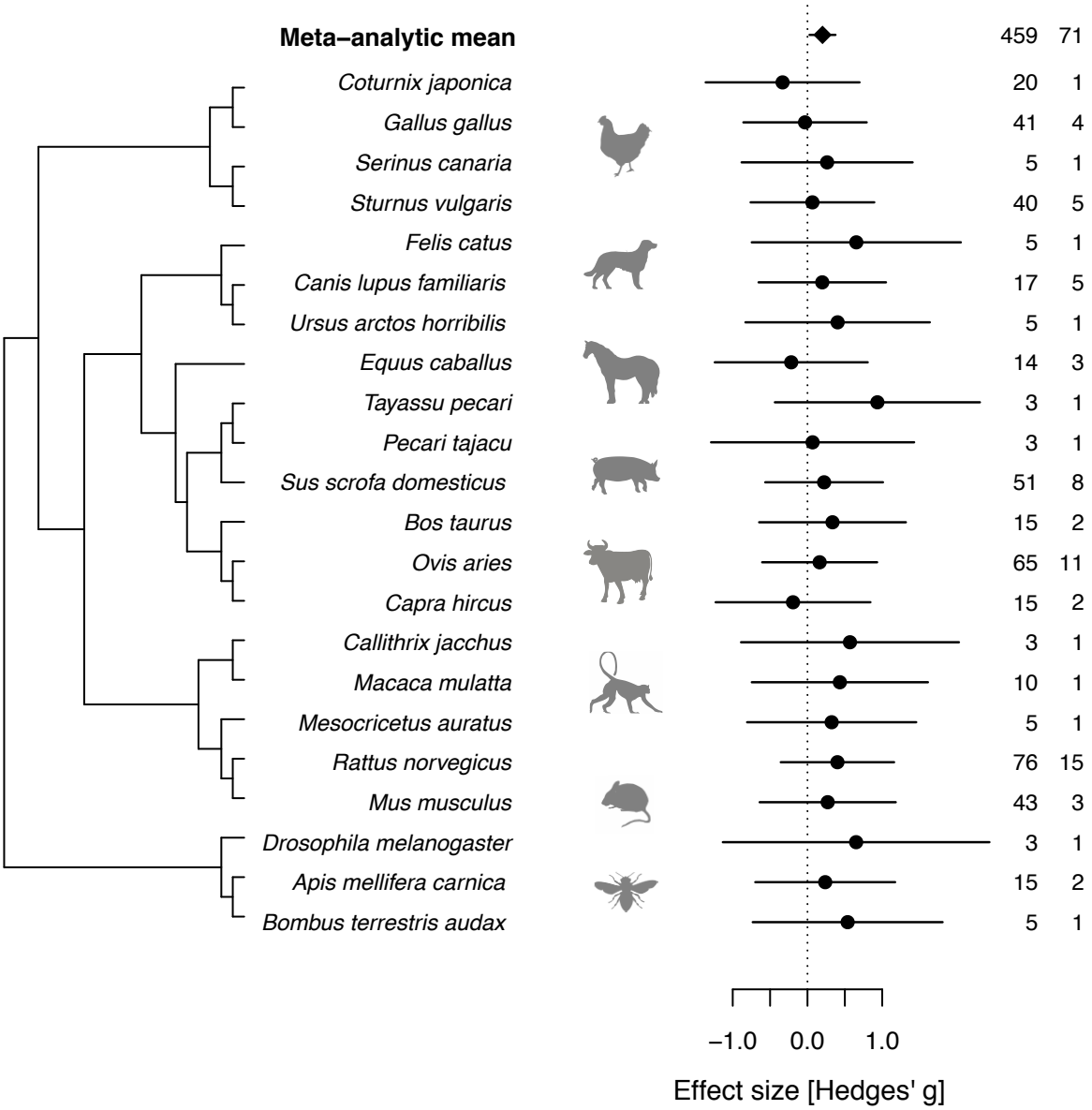
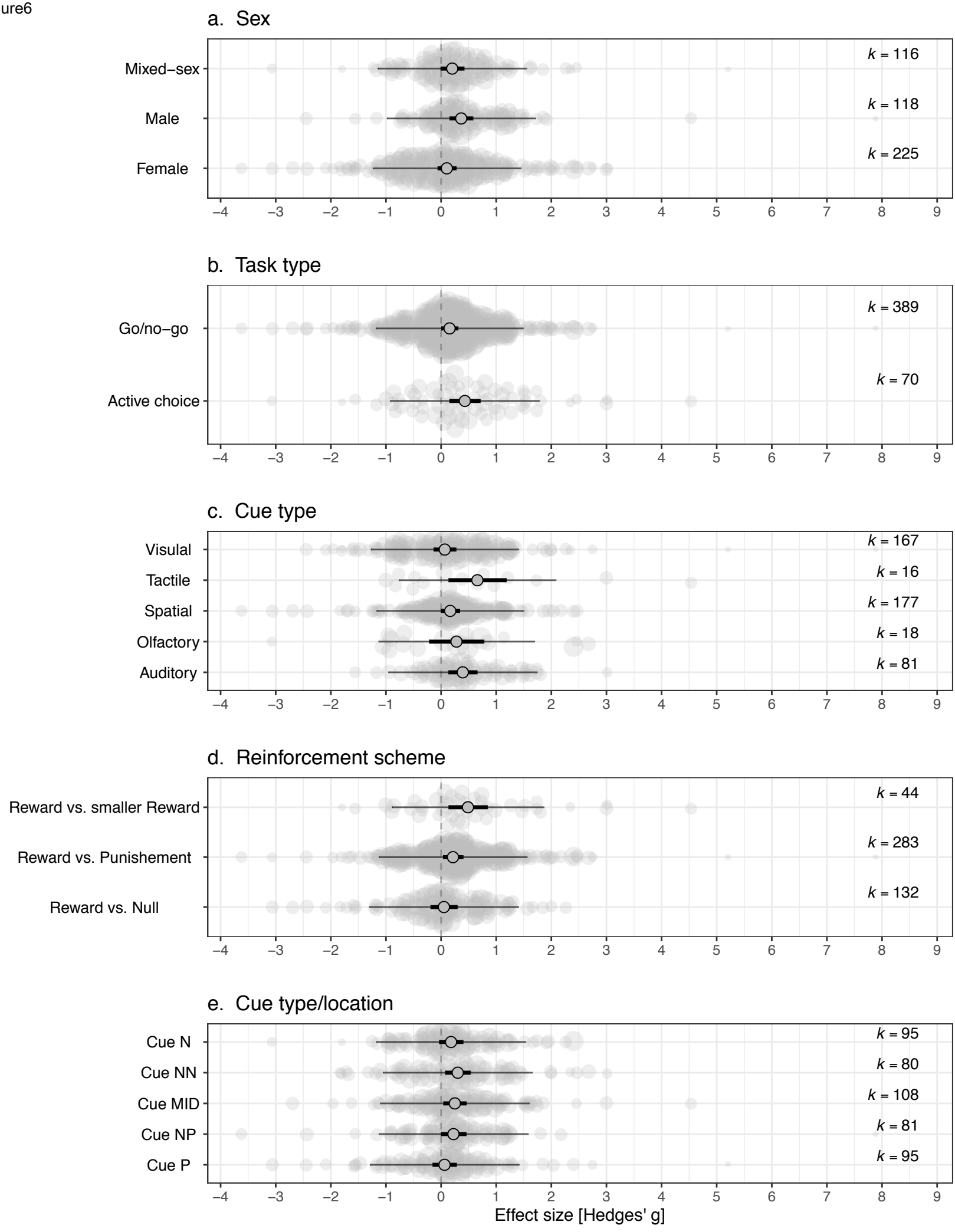

k
K
Meta–analytic mean
459 71
Coturnix japonica
20 1
Gallus gallus
41 4
Serinus canaria
5 1
Sturnus vulgaris
40 5
Felis catus
5 1
Canis lupus familiaris
17 5
Ursus arctos horribilis
5 1
Equus caballus
14 3
Tayassu pecari
3 1
Pecari tajacu
3 1
Sus scrofa domesticus
51 8
Bos taurus
15 2
Ovis aries
65 11
Capra hircus
15 2
Callithrix jacchus
3 1
Macaca mulatta
10 1
Mesocricetus auratus
5 1
Rattus norvegicus
76 15
Mus musculus
43 3
Drosophila melanogaster
3 1
Apis mellifera carnica
15 2
Bombus terrestris audax
5 1
–1.0
0.0
1.0
Effect size [Hedges' g]

Figure6


a. Sex
Mixed−sex
Male
Female
k = 116
k = 118
k = 225
b. Task type
Go/no−go
Active choice
k = 389
k = 70
c. Cue type
Visulal
Tactile
Spatial
Olfactory
Auditory
k = 167
k = 16
k = 177
k = 18
k = 81
d. Reinforcement scheme
Reward vs. smaller Reward
Reward vs. Punishement
Reward vs. Null
k = 44
k = 283
k = 132
e. Cue type/location
Cue N
Cue NN
Cue MID
Cue NP
Cue P
k = 95
k = 80
k = 108
k = 81
k = 95
−4 −3 −2 −1 0 1 2 3 4 5 6 7 8 9
Effect size [Hedges' g]

Appendix A

Optimism, pessimism and judgement bias in animals:

a systematic review and a meta-analysis

Lagisz, Malgorzata [1†], Zidar, Josefina[2,†], Nakagawa, Shinichi[1,†*], Neville, Vikki[3], Soratro, Enrico[2], Paul, Elizabeth S.[3], Bateson, Melissa[4], Mendl, Michael[3#], Løvlie, Hanne[2#]

Addresses:

[1] Evolution and Ecology Research Centre, School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, New South Wales, Sydney, NSW 2052, Australia

[2] The Department of Physics, Chemistry and Biology, IFM Biology, Linköping University, SE-581 83 Linköping, Sweden

[3] Centre for Behavioural Biology, Bristol Veterinary School, University of Bristol, Langford, BS40 5DU, United Kingdom

[4] Centre for Behaviour and Evolution, Biosciences Institute, Newcastle University, Newcastle upon Tyne, NE2 4HH, United Kingdom

*__Correspondence__: Shinichi Nakagawa s.nakagawa@unsw.edu.au

[†] These authors contributed equally to this work

[#] These authors supervised this work equally and are joint senior authors

Data and code available at OSF: https://osf.io/anfhm/

Supplementary Methods

Decision tree for the classification of treatments as inducing 'relatively better' or 'relatively worse' affective states is presented in Figure S1. The overview of the effect size extraction and calculations are presented as a diagram in Figure S2. In the original papers, measurements of animal behaviour during judgement bias trials were reported either as latencies (time to task/outcome) or proportions/percentages of trials where given outcome/action was achieved during specified time period. Because latency and proportion data are bounded (i.e. latencies start at 0 and are often censored, and proportions are between 0-100), we used natural log(latencies) or logit(proportions) transformed data to calculate our effect sizes, Hedges' *g*. To calculate Hedges' *g*, we subtracted the ln-transformed or logit-transformed mean value of the relatively worse treatment from the ln- transformed or logit-transformed mean of the relatively better treatment, and divided the difference by the pooled SD with correction for small sample sizes (Hedges & Olkin, 1985). The formulas used to calculate effect sizes and their variances are shown in Figure S2.

As indicated in the main text, to be able to compare latency and proportion data (e.g. short latencies to go to an ambiguous location are interpreted as an 'optimistic' response, whilst a high proportion of decisions to visit the ambiguous locations are also interpreted as 'optimistic'), we reversed the sign of Hedges' *g* for the latency data. Thus, after reversing Hedges' *g* for latency data, all positive effect sizes can be interpreted as optimistic responses of animals in relatively better affective states compared to those in relatively worse affective states.

Supplementary Results

Funnel plots for publication bias assessment is presented in Figure S4. Table S1 contains detailed descriptions of the extracted data (meta-data), while Tables S2 and S3 present lists of included and excluded (at full-text screening stage) papers, respectively. Tables S4 to S22 show results of meta-analytic and meta-regression models for the full data set and data subsets. Tables S23 to S26 show results of sensitivity analyses.

Supplementary References

Hedges, L., Olkin, I. (1985). Statistical methods for meta-analysis. Academic Press, New York.

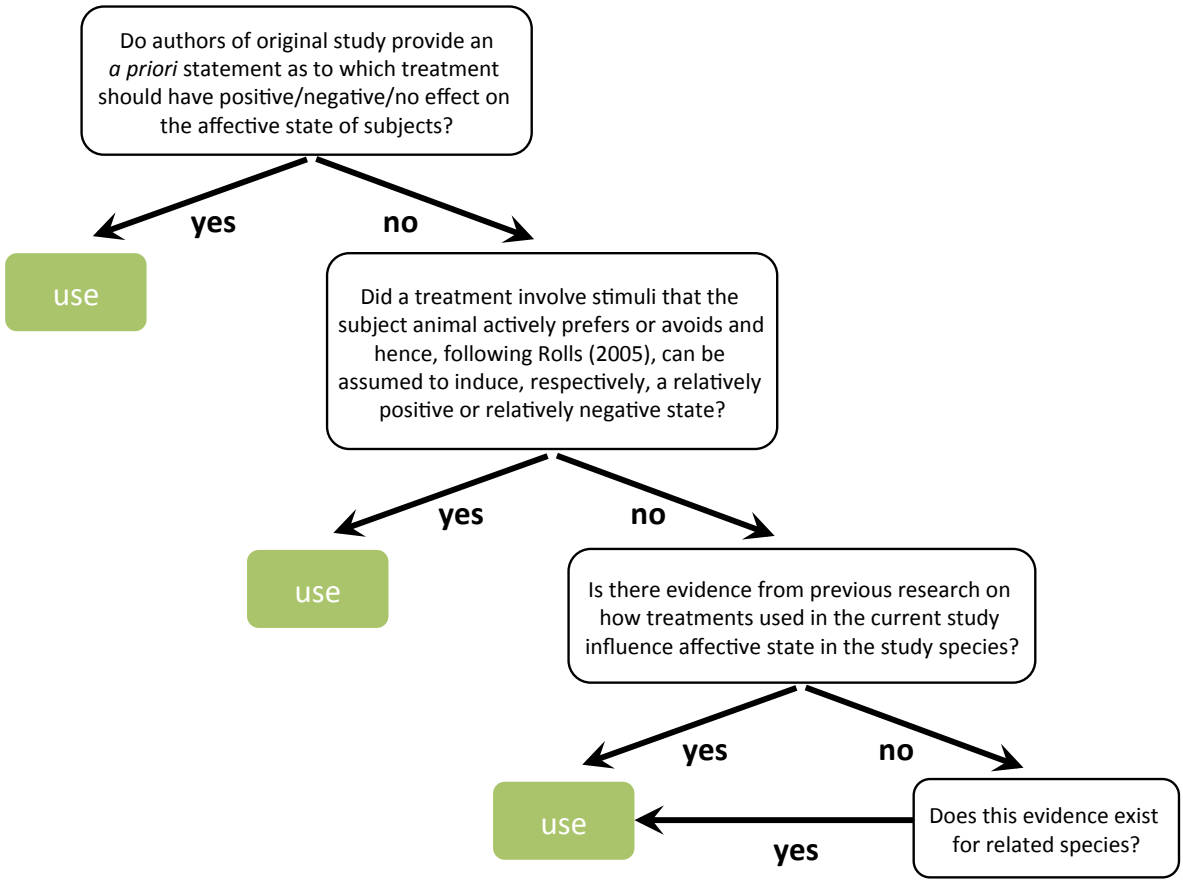


Figure S1

Decision tree for classification of treatments within a study as inducing 'relatively better' or 'relatively worse' affective states.

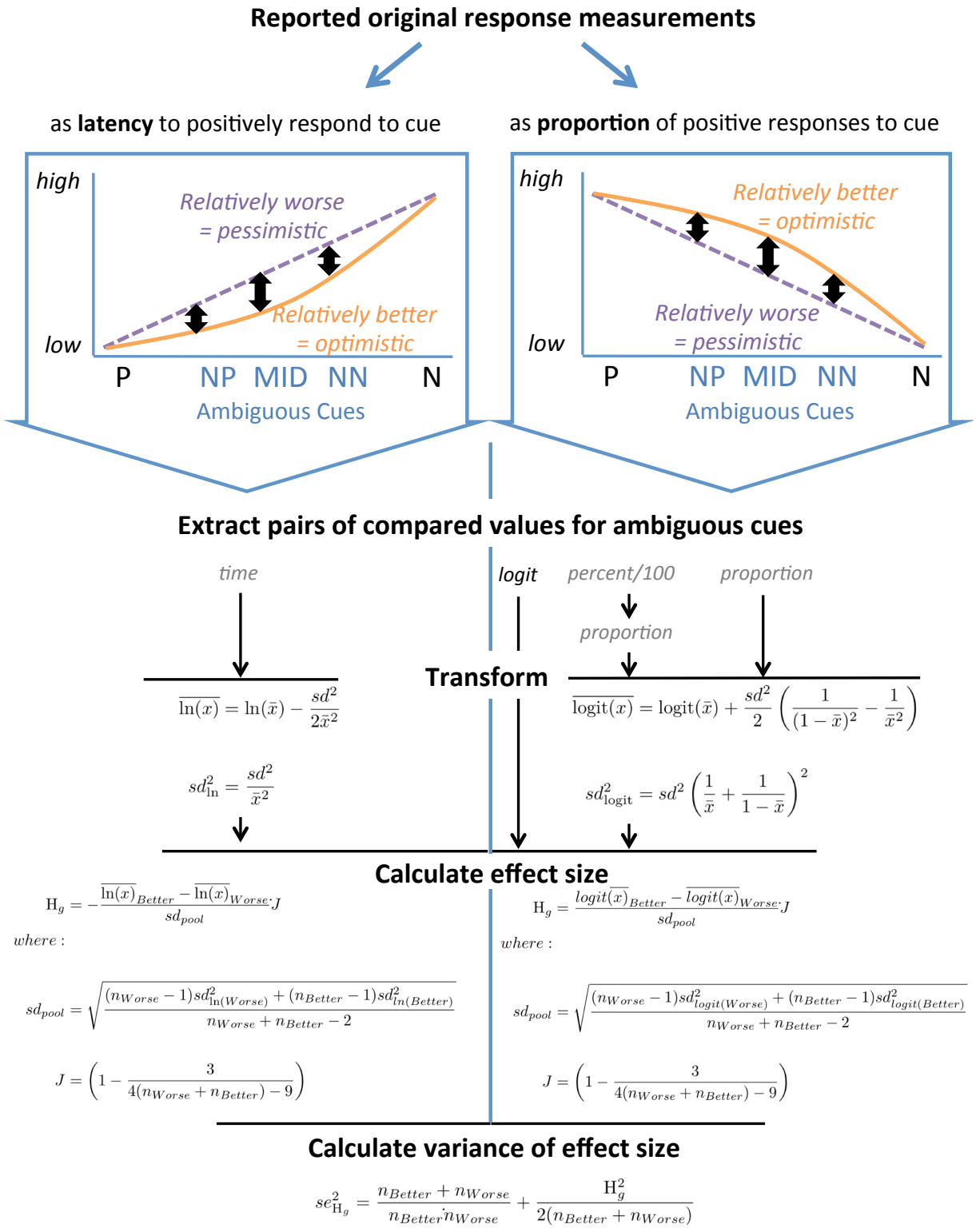
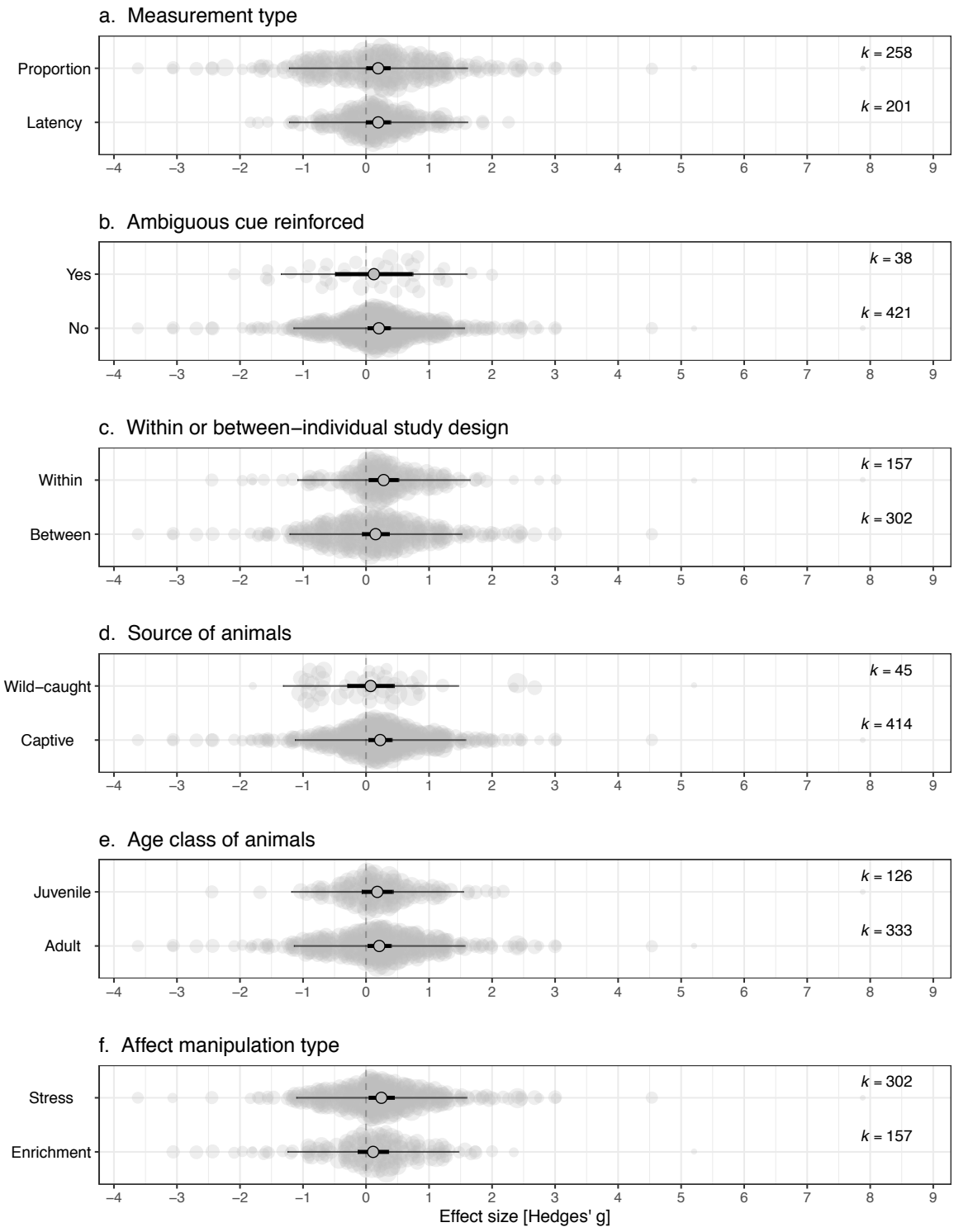


Figure S2

Workflow of data extraction, transformation and calculation of effect sizes.



(figure continued on next page)

(figure continued from previous page)

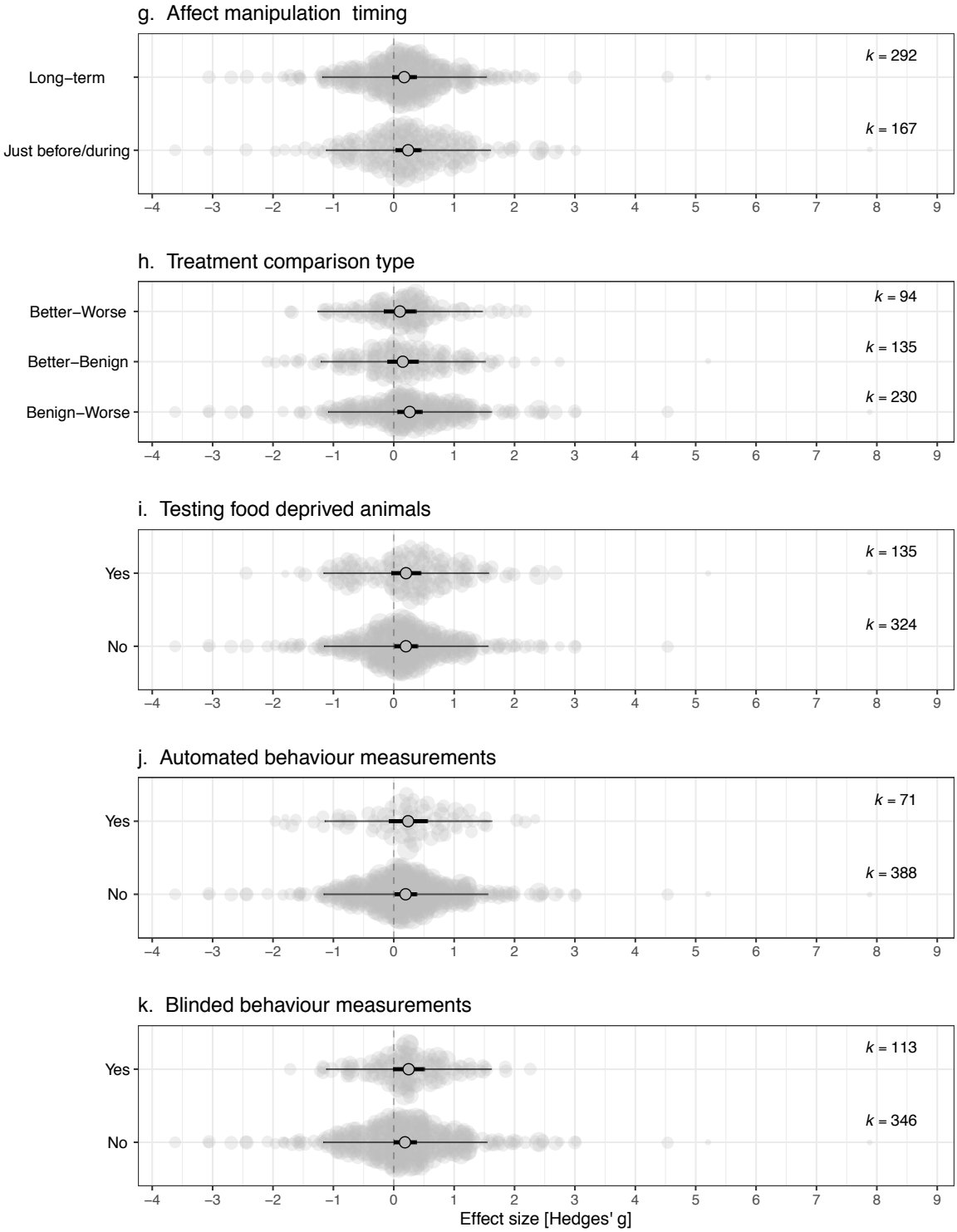


Figure S3

Forest plots showing effect size (Hedges' g) estimates from the univariate meta-regression analyses (one moderator at a time) with potentially non-influential moderators. Effect sizes with positive values indicate a positive effect of affect manipulations on judgement bias in a relatively better condition compared to a relatively worse condition, i.e. affect manipulation treatment working in the expected direction. The mean effects (black unfilled circles) for each group of individual effect sizes (grey filled circles) are statistically different from zero when their horizontal error bars (95% confidence intervals) do not cross zero. Thin horizontal whiskers indicate prediction intervals. *k* is number of effect sizes. Dots represent individual effect sizes scaled proportionally to their precision.

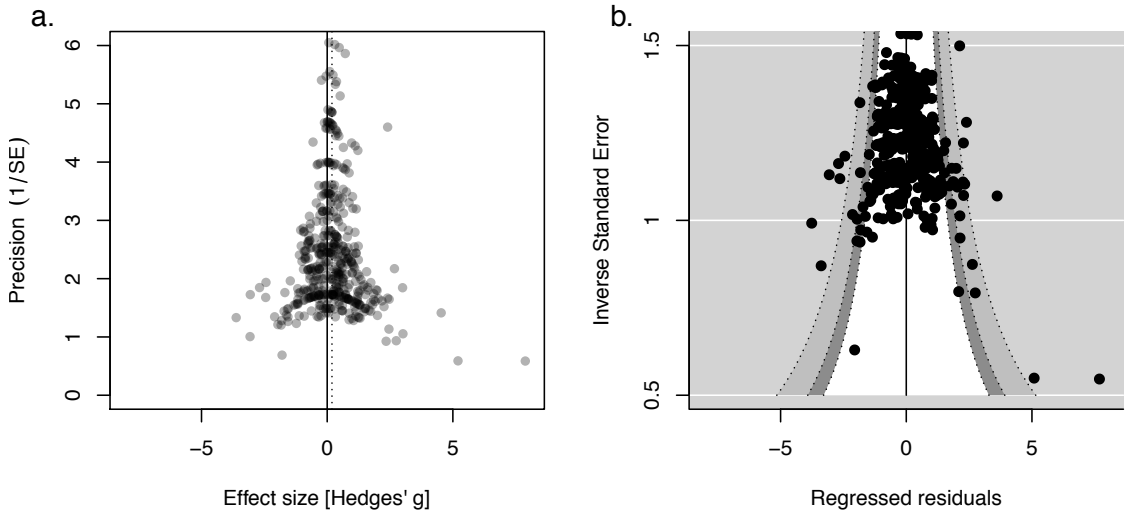


Figure S4

Funnel plots for the estimated effect sizes for meta-analytic (intercept-only) models for the difference between better and worse treatments (Hedges' g, *Hg*). a) Raw effect sizes and their precision (inversed standard errors); vertical solid lines indicate zero (i.e. where there is no effect of the affect manipulations) and the dashed line the meta-analytic mean. b) The residual values from the full meta-regression models, containing animals' sex, task type, cue type and reinforcement scheme used in judgement bias tests, plotted against inversed standard errors. Positive values of *Hg* can be interpreted as animals in relatively better treatment group being more 'optimistic' than animals in the relatively worse treatment group, i.e. affect manipulation treatment working in the expected direction.

Table S1

Column names and descriptions of the contents of each column in the unprocessed data extracted from the original experimental papers. For the definition of Better/Worse groups see the main text, main text Methods section, and Figure S1.

Column name	Column description and code values
EffectID	Unique ID for each pairwise comparison used to calculate effect sizes
ArticleID	Unique ID for each extracted original experimental paper
ExperimentID	Unique ID for each extracted experiment within the original paper
GroupID	Unique ID for each group of animals used in experiment within the original paper
Authors	Authors of the original paper
Journal	Publication journal of the original paper
JournalType	Type of journal: • peer-reviewed = any academic journal • unpublished = data and descriptions from the authors only
Year	Publication year of the original paper
ArticleTitle	Title of the original paper
Species	Common name of an animal species used in the experiment
ScientificName	Scientific name of an animal species used in the experiment
Taxa	High-level taxon of the species: • mammal = species from class Mammalia • bird = species from class Aves • insect = species from class Insecta
Breed	Breed of an animal species used in the experiment, for domesticated animals, as eported by the study authors, otherwise (for non-domesticated animals) coded as not applicable [n/a]
Captive_Wild-caught	Source of animals used in the experiment, as reported in the paper: • captive = all used animals were captive, or source not reported • wild-caught = all used animals were wild-caught
Age	Age of animals used in the experiment: • juvenile = all used animals were not sexually mature • adult = all used animals were sexually mature, mixed age, or age not reported
WithinBetween	Whether between-individual or within-individual study design was used: • between = two or more groups of animals were simultaneously subject to different treatments (or treatment vs. control/benign) • within = same group of animals was subject sequentionally to different treatments (or treatment vs. control/beningn), includes cross-over design
CrossoverDesign	Whether crossover experimental design was used, if within-subject design was used: • yes = all groups of animals received treatment in a different order, so that all animals received each treatment at some time point • no = each group of animals was subject to only one treatment (or

	control/benign) at the same time • n/a = for between-individual study design
StudyDesign	Combined information about study design from WithinBetween and CrossoverDesign columns: • within (before-after) = same group of animals was subject first to no treatment and then to treatment • within (crossover) = same group of animals was subject sequentially to different treatments (or treatment vs. control/benign) in different order • between = two or more groups of animals were simultaneously subject to different treatments (or treatment vs. control/benign)
Blind	Whether testing animals for judgement bias was blinded, as reported by the study authors: • yes = authors of the original study explicitly state that researchers performing trials were blinded to the treatment groups of the tested animals • no = authors state that state that researchers performing trials were not blinded to the treatment groups of the tested animals, or blinding was not possible. If not info was provided on blinding being used, we assumed no blinding
Automated	Whether measuring animal behaviour during trials was automated: • yes = authors explicitly state that the trials/measurements were automated • no = the methodological description of the experiment does not include any mention of automation of the testing process
FoodDeprived	Whether animals were food-deprived before the behavioural trials: • yes = authors explicitly state that the animals were food-deprived before the trials • no = the methodological description of the experiment does not include any mention that the animals were food restricted before the trials
TaskType	Type of the task used during behavioural trials: • active choice = go/go tasks in which an animal is required to make an active response to cues perceived as positive and to cues perceived as negative • go/no-go = tasks in which an animal is required to suppress a response to cues perceived as negative and actively respond only to cues perceived as positive
CueTypeDetails	Type of the cue used during training and trials: • auditory = any sound-based cues • colour = places, objects, shapes differing only in colour parameters • light = any light-based cues, but not related to colour or location (e.g. light on/off) • location = different places in physical space (e.g. left vs. right side, coded as spatial in CueTypeCat) • odour = any odour-based cues • shape = objects differing only in shape • tactile = any tactile-based cues • visual = any other visual cues not covered by other categories above
CueTypeCat	Alternative, simplified, classification of type of the cue used: • auditory = any sound-based cues • spatial = different places in physical space, e.g. left vs. right side (usually different locations within the test chamber) • olfactory = any odour-based cues • tactile = any tactile-based cues • visual = any visual cues

ResponseTypeDetails	Details of what type of animal response was measured as an outcome
OutcomeCorrectPosCue	Details of an outcome for an animal correctly responding to positive cue
OutcomeIncorrectPosCue	Details of an outcome for an animal incorrectly responding to positive cue
OutcomeCorrectNegCue	Details of an outcome for an animal correctly responding to negative cue
OutcomeIncorrectNegCue	Details of an outcome for an animal incorrectly responding to negative cue
ReinforcementCat	How reinforcement (R = reward; P = punishment; Null = no reward/punishment) was used during training: • R-Null = reward for the positive cue and nothing for the negative cue • R-P = reward for the positive cue and punishment for the negative cue • R-R = larger reward for the positive cue and smaller reward for the negative cue
AffectManipDetails	Brief details of affect manipulation (what was actually manipulated, potentially influencing affect), based on the description in the original paper
AffectManipCat	Category of affect manipulation type (manipulation potentially influencing affect): • enrichment = environmental enrichment or other positive events used as an affect manipulation treatment • stress = stress or potentially negative events used as an affect manipulation treatment
AffectManipTiming	Timing of affect manipulation: • before/during = affect manipulation was applied within hours before the judgement bias test and/or during the judgement bias test (usually in acute treatments) • long-term = affect manipulation was applied for longer than a few hours before the judgement bias test (usually in chronic treatments)
AmbigReinforced	Whether ambiguous cues were reinforced: • yes = ambiguous cues were rewarded fully, partially or randomly during the judgement bias test • no = ambiguous cues were not rewarded during the judgement bias trials. If no information provided, we assumed no reinforcement
NoAmbigCues	Number of different ambiguous cues used in the judgement bias trials
NoTestSessions	Number of test sessions (e.g. separate days) with the judgement bias trials
NoTrialsAmbigCue	Number of tests per ambiguous cue/session
NoTrialsTrainingCue	Number of tests per training cue/session during training
MeasureType	Whether outcome measure was reported as latency or proportion: • latency = outcome measure reported as a time from trial start to certain action or criteria • proportion = proportion (or percentage) of trials/tests in which animal performed certain action or met certain criteria
BetterSampleSizeMale	Number of males in the Better group, if reported; n/a if not reported
BetterSampleSizeFemale	Number of females in the Better group, if reported; n/a if not reported
WorseSampleSizeMale	Number of males in the Worse group, if reported; n/a if not reported
WorseSampleSizeFemale	Number of females in the Worse group, if reported; n/a if not reported
Sex	Sex of tested animals in the compared groups: • female = only female animals were used

	• male = only male animals were used • both = both female and male animals were used
TreatmentComp	Details of affect manipulation performed on animals, i.e. what types of groups were compared (e.g. stressed vs. control, enriched vs. barren)
ComparisonCat	Comparison category according to the direction of comparison and affect manipulation type of the compared animal groups: • Better-Worse = when one group of animals was subject to manipulation that was better (or could be expected to induce positive affect) than benign condition and the other group of animals was subject to a treatment that was worse than benign condition (or could be expected to induce negative affect) • Benign-Worse = when one group of animals was subject to a treatment considered as benign condition (no change to affect or neutral affect) and the other group of animals was subject to a treatment that was worse than benign condition (or could be expected to induce negative affect) • Better-Benign = when one group of animals was subject to a treatment that was better than benign condition (or could be expected to induce positive affect) and the other group of animals was subject to a treatment considered as benign condition (no change to affect or neutral affect)
ScalePoint	Codes of the positions of the ambiguous cues relatively to the the trained positive and negative cues: • A = 'near positive' = extracted intermediate cue between the middle (midpoint) cue and the rewarded cue, if more than 1 ambiguous cue was used • B = 'midpoint' = extracted middle cue between the trained cues • C = 'near negative' = extracted intermediate cues between the middle cue (midpoint) and the unrewarded cue, if more than 1 ambiguous cue was used
Better	Mean value of outcome for the Better group
BetterSE	Standard error of outcome for the Better group
BetterN	Sample size (number of animals tested) of outcome for the Better group
BetterSD	Standard deviation of outcome for the Better group
Worse	Mean value of outcome for the Worse group
WorseSE	Standard error of outcome for the Worse group
WorseN	Sample size (number of animals tested) of outcome for the Worse group
WorseSD	Standard deviation of outcome for the Worse group
DataScale	Whether the extracted data is on logit or natural scale
DataSE	Whether Standard Error for the outcome was reported in the original study: • yes = Standard Error for the outcome measure reported in the original paper • no = Standard Error for the outcome measure not reported in the original paper
DataSource	Details of location in the original paper (Figure, Table) from which the outcome data was extracted
Notes	Any other relevant information
Exclude	"Y" denotes any data points that were excluded after data extraction and before analyses. Reasons for exclusions are recorded in Notes

Table S2

List of included studies. Main characteristics of the included studies are summarised in Supplementary File 1, while the complete data set (as extracted) is provided on GitHub. Stars indicate studies where raw data or additional data was received from the authors upon request.

Nr	ArticleID	Authors	Year	Journal	ArticleTitle
1	Ash2016	Ash, H., Buchanan-Smith, H.M.	2016	Applied Animal Behaviour Science	The long-term impact of infant rearing background on the affective state of adult common marmosets (Callithrix jacchus)
2	Asher2016	Asher, L., Friel, M., Griffin, K., Collins, L.M.	2016	Biology Letters	Mood and personality interact to determine cognitive biases in pigs
3	Baciadonna2016	Baciadonna, L., Nawroth, C., McElligott, A.G.	2018	PeerJ	Judgement bias in goats (Capra hircus): investigating the effects of human grooming
4	Bailoo2018*	Bailoo, J.D., Murphy, E., Boada-Saña, M., Varholick, J.A., Hintze, S., Baussière, C., Hahn, K.C., Göpfert, C., Palme, R., Voelkl, B., Würbel, H.	2018	Frontiers in Behavioral Neuroscience	Effects of cage enrichment on behavior, welfare and outcome variability in female mice
5	Barker2017a*	Barker, T.H., George, R.P., Howarth, G.S., Whittaker, A.L.	2017	PLOS One	Assessment of housing density, space allocation and social hierarchy of laboratory rats on behavioural measures of welfare the absence of a physiological stress response
6	Barker2017b	Barker, T.H., Bobrovskaya, L., Howarth, G.S., Whittaker, A.L.	2017	Physiology and Behavior	Female rats display fewer optimistic responses in a judgment bias test in the absence of a physiological stress response
7	Bateson2007	Bateson, M., Matheson, S. M.	2007	Animal Welfare	Performance on a categorisation task suggests that removal of environmental enrichment induces 'pessimism' in captive European starlings (Sturnus vulgaris)
8	Bateson2011	Bateson, M., Desire, S., Gartside, S. E., Wright, G. A.	2011	Current Biology	Agitated honeybees exhibit pessimistic cognitive biases
9	Bateson2015b	Bateson, M., Emmerson, M., Ergün, G., Monaghan, P., Nettle, D.	2015	PLOS One	Opposite effects of early-life competition and developmental telomere attrition on cognitive biases in juvenile European starlings

10	Bethell&Koyama2015	Bethell, E. J., Koyama, N. F.	2015	Royal Society Open Science	Happy hamsters? Enrichment induces positive judgement bias for mildly (but not truly) ambiguous cues to reward and punishment in Mesocricetus auratus
11	Bethell2012	Bethell, E. J., Holmes, A., MacLarnon, A., Semple, S.	2012	Animal Welfare	Cognitive bias in a non-human primate: Husbandry procedures influence cognitive indicators of psychological well-being in captive rhesus macaques
12	Boleij2012	Boleij, H., Klooster, J. V., Lavrijsen, M., Kirchhoff, S., Arndt, S. S., Ohl, F.	2012	Behavioural Brain Research	A test to identify judgement bias in mice
13	Brajon2015*	Brajon, S., Laforest, J. P., Schmitt, O., Devillers, N.	2015	PLOS One	The way humans behave modulates the emotional state of piglets
14	Briefer2013*	Briefer, E. F., McElligott, A. G.	2013	Applied Animal Behaviour Science	Rescued goats at a sanctuary display positive mood after former neglect
15	BrieferFreymond2014	Briefer Freymond, S., Briefer, E. F., Zollinger, A., Gindrat-von Allmen, Y., Wyss, C., Bachmann, I.	2014	Applied Animal Behaviour Science	Behaviour of horses in a judgment bias test associated with positive or negative reinforcement
16	Brilot2010*	Brilot, B. O., Asher, L. , Bateson, M.	2010	Animal Cognition	Stereotyping starlings are more 'pessimistic'
17	Brydges2011	Brydges, N. M., Leach, M., Nicol, K., Wright, R., Bateson, M.	2011	Animal Behaviour	Environmental enrichment induces optimistic cognitive bias in rats
18	Brydges2012	Brydges, N. M., Hall, L., Nicolson, R., Holmes, M. C., Hall, J.	2012	PLOS One	The Effects of Juvenile Stress on Anxiety, Cognitive Bias and Decision Making in Adulthood: A Rat Model
19	Burman2008	Burman, O. H. P., Parker, R., Paul, E. S., Mendl, M.	2008	Animal Behaviour	A spatial judgement task to determine background emotional state in laboratory rats, Rattus norvegicus
20	Burman2009*	Burman, O. H. P., Parker, R. M. A., Paul, E. S., Mendl, M. T.	2009	Physiology and Behavior	Anxiety-induced cognitive bias in non-human animals
21	Burman2011	Burman, O., McGowan, R., Mendl, M., Norling, Y., Paul, E., Rehn, T., Keeling, L.	2011	Applied Animal Behaviour Science	Using judgement bias to measure positive affective state in dogs
22	Carreras2016b	Carreras, R., Mainau, E., Arroyo, L.,	2016	Applied Animal Behaviour	Housing conditions do not alter cognitive bias but affect serum cortisol, qualitative behaviour assessment and

		Moles, X., Gonzales, J., Bassols, A., Dalmau, A., Faucitano, L., Manteca, X., Velarde, A.		Science	wounds on the carcass in pigs
23	Carreras2017	Carreras, R., Arroyo, L., Mainau, E., Valent, D., Bassols, A., Dalmau, A., Faucitano, L., Manteca, X., Velarde, A.	2017	Behavioural Process	Can the way pigs are handled alter behavioural and physiological measures of affective state?
24	Coulon2015	Coulon, M., Nowak, R., Andanson, S., Petit, B., Lévy, F., Boissy, A.	2015	Developmental Psychobiology	Effects of prenatal stress and emotional reactivity of the mother on emotional and cognitive abilities in lambs
25	Daros2014	Daros, R. R., Costa, J. H. C., von Keyserlingk, M. A. G., Hoetzel, M. J., Weary, D. M.	2014	PLOS One	Separation from the dam causes negative judgement bias in dairy calves
26	Deakin2016	Deakin, A., Browne, W. J., Hodge, J. J. L., Paul, E. S., Mendl, M.	2016	PLOS One	A Screen-Peck Task for Investigating Cognitive Bias in Laying Hens
27	Deakin2018	Deakin, A., Mendl, M., Browne, W. J., Paul, E. S., Hodge, J. J. L.	2018	Biology Letters	State-dependent judgement bias in Drosophila: evidence for evolutionarily primitive affective processes
28	Destrez2013	Destrez, A., Deiss, V., Lévy, F., Calandreau, L., Lee, C., Chaillou-Sagon, E., Boissy, A.	2013	Applied Animal Behaviour Science	Chronic stress induces pessimistic-like judgment and learning deficits in sheep
29	Destrez2014	Destrez, A., Deiss, V., Leterrier, C., Calandreau, L., Boissy, A.	2014	Applied Animal Behaviour Science	Repeated exposure to positive events induces optimistic-like judgment and enhances fearfulness in chronically stressed sheep
30	Destrez2017	Destrez, A., Boissy, A., Guilloteau, L., Andanson, S., Souriau, A., Laroucau, K., Chaillou, E., Deiss, V.	2017	Animal	Effects of a chronic stress treatment on vaccinal response in lambs
31	Douglas2012	Douglas, C., Bateson, M., Walsh, C., Bédué, A., Edwards, A.	2012	Applied Animal Behaviour Science	Environmental enrichment induces optimistic cognitive biases in pigs

		S.			
32	Doyle2010a	Doyle, R. E., Fisher, A. D., Hinch, G. N., Boissy, A., Lee, C.	2010	Applied Animal Behaviour Science	Release from restraint generates a positive judgement bias in sheep
33	Doyle2011b	Doyle, R. E., Lee, C., Deiss, V., Fisher, A. D., Hinch, G. N., Boissy, A.	2011	Physiology and Behavior	Measuring judgement bias and emotional reactivity in sheep following long-term exposure to unpredictable and aversive events
34	Dupjan2013	Dûpjan, S., Ramp, C., Kanitz, E., Tuchscherer, A., Puppe, B.	2013	Journal of Veterinary Behavior: Clinical Applications and Research	A design for studies on cognitive bias in the domestic pig
35	Duranton2019	Duranton, C., Horowitz, A.	2019	Applied Animal Behaviour Science	Let me sniff! Nosework induces positive judgment bias in pet dogs
36	Gott2019	Gott, A., Andrews, C., Bedford, T., Nettle, D., Bateson, M.	2019	Animal Cognition	Developmental history and stress responsiveness are related to response inhibition, but not judgement bias, in a cohort of European starlings (Sturnus vulgaris)
37	Guldimann2015	Guldimann, K., Vögeli, S., Wolf, M., Wechsler B., Gygax, L.	2015	Brain and Cognition	Frontal brain deactivation during a non-verbal cognitive judgement bias test in sheep
38	Hales2016	Hales, C. A., Robinson, E. S. J., Houghton, C. J.	2017	PLOS One	Diffusion Modelling Reveals the Decision Making Processes Underlying Negative Judgement Bias in Rats
39	Harding2004	Harding, E. J., Paul, E., Mendl, M.	2004	Nature Brief Communications	Cognitive bias and affective state
40	Horváth2016	Horváth, M., Pichová, K., Koˇst’ál, L.	2016	Applied Animal Behaviour Science	The effects of housing conditions on judgement bias in Japanese quail
41	Jones2018	Jones, S., Neville, V., Higgs, L., Paul, E. S., Dayan, P., Robinson, E. S. J., Mendl, M.	2018	Scientific Reports	Assessing animal affect: an automated and self-initiated judgement bias task based on natural investigative behaviour
42	Kasbaoui2016	Kasbaoui, N., Cooper, J., Mills, D. S., Burman, O.	2016	PLOS One	Effects of Long-Term Exposure to an Electronic Containment System on the Behaviour and Welfare of Domestic Cats
43	Keen2014	Keen, H. A., Nelson, O. L., Robbins, C. T., Evans, M., Shepherdson, D.	2014	Animal Cognition	Validation of a novel cognitive bias task based on difference in quantity of reinforcement for assessing environmental enrichment

		J., Newberry, R. C.			
44	Lalot2017	Lalot, M., Ung, D., Péron, F., d'Ettorre, P., Bovet, D.	2017	Behavioural Process	You know what? I'm happy. Cognitive bias is not related to personality but is induced by pair-housing in canaries (Serinus canaria)
45	Löckener2015*	Löckener, S., Reese, S., Erhard, M., Wöhr, A-C.	2015	journal of Veterinary Behavior: Clinical Applications and Research	Pasturing in herds after housing in horseboxes induces a positive cognitive bias in horses
46	Matheson2008*	Matheson, S. M., Asher, L., Bateson, M.	2008	Applied Animal Behaviour Science	Larger, enriched cages are associated with 'optimistic' response biases in captive European starlings (Sturnus vulgaris)
47	McGuire2018	McGuire, M. C., Johnson-Ulrich, Z., Robeson, A., Zeigler-Hill, V., Vonk, J.	2018	Journal of Veterinary Bahavior	I say thee "neigh": Rescued equids are optimistic in a judgment bias test
48	Mueller2012*	Mueller, C. A., Riemer, S. , Rosam, C. M., Schoesswender, J., Range, F., Huber, L.	2012	Animal Cognition	Brief owner absence does not induce negative judgement bias in pet dogs
49	Murphy2013*	Murphy, E., Nordquist, R. E., van der Staay, F. J.	2013	Applied Animal Behaviour Science	Responses of conventional pigs and Gottingen miniature pigs in an active choice judgement bias task
50	Neave2013	Neave, H. W., Daros, R. R., Costa, J. H. C., Von Keyserlingk, M. A. G., Weary, D. M.	2013	PLOS One	Pain and pessimism: Dairy calves exhibit negative judgement bias following hot-iron disbudding
51	Nogueira da Cunha2015	Nogueira da Cunha, S. Siqueira, F., Iurianny, K., Oliveira, C., Thaise, S. G., Nogueira-Filho, S. L. G., Mendl, M.	2015	PLOS One	Does Trapping Influence Decision-Making under Ambiguity in White-Lipped Peccary (Tayassu pecari)?
52	Novak2016	Novak, J., Stojanovski, K., Melotti, L., Reichlin, T.S., Palme, R., Würbel, H.	2016	Applied Animal Behaviour Science	Effects of stereotypic behaviour and chronic mild stress on judgement bias in laboratory mice
53	Oliveira2016	Oliveira, F. R. M., Nogueira-Filho, S. L. G., Sousa, M. B. C., Dias, C. T. S, Mendl, M., Nogueira, S. S. C.	2016	Applied Animal Behaviour Science	Measurement of cognitive bias and cortisol levels to evaluate the effects of space restriction on captive collared peccary (Mammalia, Tayassuidae)

54	Papciak2013	Papciak, J., Popik, P., Fuchs, E., Rygula, R.	2013	Behavioural Brain Research	Chronic psychosocial stress makes rats more 'pessimistic' in the ambiguous-cue interpretation paradigm
55	Parker2014	Parker, R. M. A., Paul, E. S., Burman, O. H. P., Browne, W. J., Mendl, M.	2014	Behavioural Brain Research	Housing conditions affect rat responses to two types of ambiguity in a reward-reward discrimination cognitive bias task
56	Perry2016	Perry, C. J., Baciadonna, L., Chitka, L.	2016	Science	Unexpected rewards induce dopamine-dependent positive emotion–like state changes in bumblebees
57	Richter2012*	Richter, S. H., Schick, A., Hoyer, C., Lankisch, K., Gass, P., Vollmayr, B.	2012	Cognitive, Affective and Behavioral Neuroscience	A glass full of optimism: enrichment effects on cognitive bias in a rat model of depression
58	Rygula2012	Rygula, R., Pluta, H., Popik, P.	2012	PLOS One	Laughing Rats Are Optimistic
59	Rygula2013	Rygula, R., Papciak, J., Popik, P.	2013	Neuropsychopharmacology	Trait pessimism predicts vulnerability to stress-induced anhedonia in rats.
60	Sanger2011	Sanger, M. E., Doyle, R. E., Hinch, G. N., Lee, C.	2011	Applied Animal Behaviour Science	Sheep exhibit a positive judgement bias and stress-induced hyperthermia following shearing
61	Schlüns2016	Schlüns, H., Welling, H., Federici, J.R., Lewejohann, L.	2017	Animal Cognition	The glass is not yet half empty: agitation but not Varroa treatment causes cognitive bias in honey bees
62	Scollo2014	Scollo, A., Gottardo, F., Contiero, B., Edwards, S. A.	2014	Applied Animal Behaviour Science	Does stocking density modify affective state in pigs as assessed by cognitive bias, behavioural and physiological parameters?
63	Seehuus2013*	Seehuus, B., Mendl, M., Keeling, L. J., Blokhuis, H.	2013	Applied Animal Behaviour Science	Disrupting motivational sequences in chicks: Are there affective consequences?
64	Uccheddu2018	Uccheddu, S., Mariti, C., Sannen, A., Vervaecke, H., Arnout, H., Rufo, Gutierrez, J., Gazzano, A., Haverbeke, A.	2018	Dog Behavior	Behavioral and cortisol responses of shelter dogs to a cognitive bias test after olfactory enrichment with essential oils
65	Verbeek2014a	Verbeek, E., Ferguson, D., Lee, C.	2014	Physiology and Behavior	Are hungry sheep more pessimistic? The effects of food restriction on cognitive bias and the involvement of ghrelin in its regulation
66	Verbeek2019	Verbeek, E., Colditz, I., Blache, D., Lee, C.	2019	PLOS One	Chronic stress influences attentional and judgement bias and the activity of the HPA axis in sheep

67	Voegeli2014*	Voegeli, S., Lutz, J., Wolf, M., Wechsler, B., Gygax, L.	2014	Behavioural Brain Research	Valence of physical stimuli, not housing conditions, affects behaviour and frontal cortical brain activity in sheep
68	Walker2014	Walker, J. K., Waran, N. K., Phillips, C. J. C.	2014	Applied Animal Behaviour Science	The effect of conspecific removal on the behaviour and physiology of pair-housed shelter dogs
69	Wheeler2015	Wheeler, R. R., Swan, M. P., Hickman, D. L.	2015	Laboratory Animals	Effect of multilevel laboratory rat caging system on the well-being of the singly-housed Sprague Dawley rat
70	Wichman2012	Wichman, A., Keeling, L. J., Forkman, B.	2012	Applied Animal Behaviour Science	Cognitive bias and anticipatory behaviour of laying hens housed in basic and enriched pens
71	Zidar 2018*	Zidar, J., Campderrich, I., Janson, E., Whichman, A., Winberg, S., Keeling, L., Løvlie, H.	2018	Scientific Reports	Environmental complexity buffers against stress-induced negative judgement bias in female chickens

Table S3

List of studies excluded during full-text screening stage, with reasons for exclusion.

Nr	Reference	Main reason for exclusion
1	Anderson, M.H., MunafÃ, M.R., Robinson, E.S.J. (2013) Investigating the psychopharmacology of cognitive affective bias in rats using an affective tone discrimination task. Psychopharmacology 226: 601–613	no affect manipulation or drug-based manipulation
2	Barker, T.H., Howarth, G.S., Whittaker, A.L. (2016) The effects of metabolic cage housing and sex on cognitive bias expression in rats. Applied Animal Behaviour Science 177: 70–76	data missing or not extractable
3	Bellegarde, L.G.A., Haskell, M.J., Duvaux-Ponter, C., Weiss, A., Boissy, A., Erhard, H.W. (2017) Face-based perception of emotions in dairy goats. Applied Animal Behaviour Science 193: 51–59	non-standard judgement bias test
4	Brilot, B.O., Bateson, M. (2012) Water bathing alters threat perception in starlings. Biology Letters 8: 379–381	non-standard judgement bias test
5	Brilot, B.O., Normandale, C.L., Parkin, A., Bateson, M. (2009) Can we use starlings' aversion to eyespots as the basis for a novel 'cognitive bias' task? Applied Animal Behaviour Science 118: 182–190	non-standard judgement bias test
6	Brydges, N.M., Hall, L. (2017) A shortened protocol for assessing cognitive bias in rats. J. Neurosci. Methods 286: 1–5	no affect manipulation or drug-based manipulation
7	Carreras, R., Arroyo, L., Mainau, E., PeÃ±a, R., Bassols, A., Dalmau, A., Faucitano, L., Manteca, X., Velarde, A. (2016) Effect of gender and halothane genotype on cognitive bias and its relationship with fear in pigs. Applied Animal Behaviour Science 177: 41973	no affect manipulation or drug-based manipulation
8	Carreras, R., Mainau, E., Rodriguez, P., Llonch, P., Dalmau, A., Manteca, X., Velarde, A. (2015) Cognitive bias in pigs: Individual classification and consistency over time. Journal of Veterinary Behavior: Clinical Applications and Research 10: 577–581	no affect manipulationor drug-based manipulation
9	Chaby, L.E., Cavigelli, S.A., White, A., Wang, K., Braithwaite, V.A. (2013) Long-term changes in cognitive bias and coping response as a result of chronic unpredictable stress during adolescence. Frontiers in Human Neuroscience 7: 328	data missing or not extractable
10	Curzytek, K., Kubera, M., Trojan, E., Wójcik, K., Basta-Kaim, A., Detka, J., Maes, M., Rygula, R. (2018) The effects of pessimism on cell-mediated immunity in rats. Progress in Neuro-Psychopharmacology and Biological Psychiatry 80: 295–303	no affect manipulation or drug-based manipulation

11	Destrez, A., Boissy, A., Guilloteau, L., Andanson, S., Souriau, A., Laroucau, K., Chaillou, E., Deiss, V. (2017) Effects of a chronic stress treatment on vaccinal response in lambs. Animal 11: 872–880	data missing or not extractable
12	Destrez, A., Deiss, V., Belzung, C., Lee, C., Boissy, A. (2012) Does reduction of fearfulness tend to reduce pessimistic-like judgment in lambs? Applied Animal Behaviour Science 139: 233–241	no affect manipulation or drug-based manipulation
13	Doyle, R.E., Hinch, G.N., Fisher, A.D., Boissy, A., Henshall, J.M., Lee, C. (2011) Administration of serotonin inhibitor p-Chlorophenylalanine induces pessimistic-like judgement bias in sheep. Psychoneuroendocrinology 36: 279–288	no affect manipulation or drug-based manipulation
14	Doyle, R.E., Lee, C., McGill, D.M., Mendl, M. (2015) Evaluating pharmacological models of high and low anxiety in sheep. PeerJ 3: e1510	no affect manipulation or drug-based manipulation
15	Doyle, R.E., Vidal, S., Hinch, G.N., Fisher, A.D., Boissy, A., Lee, C. (2010) The effect of repeated testing on judgement biases in sheep. Behavioural Processes 83: 349–352	no affect manipulation or drug-based manipulation
16	Drozd, R., Cieslak, P.E., Rychlik, M., Parkitna, J.R., Rygula, R. (2016) Cognitive judgment bias interacts with risk-based decision making and sensitivity to dopaminergic challenge in male rats. Frontiers in Behavioral Neuroscience 10: 163	no affect manipulation or drug-based manipulation
17	Drozd,. R., Rychlik, M., Fijalkowska, A., Rygula, R. (2019) Effects of cognitive judgement bias and acute antidepressant treatment on sensitivity to feedback and cognitive flexibility in the rat version of the probabilistic reversal-learning test. Behavioural Brain Research 359: 619–629	no affect manipulation or drug-based manipulation
18	Düpjan, S., Stracke, J., Tuchscherer, A., Puppe, B. (2017) An improved design for the spatial judgement task in domestic pigs. Applied Animal Behaviour Science 187: 23-30	no affect manipulation or drug-based manipulation
19	Enkel, T., Gholizadeh, D., Von Bohlen Und Halbach, O., Sanchis-Segura, C., Hurlemann, R., Spanagel, R., Gass, P., Vollmayr, B. (2010) Ambiguous-cue interpretation is biased under stress-and depression-like states in rats. Neuropsychopharmacology 35: 1008–1015	no affect manipulation or drug-based manipulation
20	Gordon, D.J., Rogers, L.J. (2015) Cognitive bias, hand preference and welfare of common marmosets. Behavioural Brain Research 287: 100–108	no affect manipulation or drug-based manipulation
21	Graulich, D.M., Kaiser, S., Sachser, N., Richter, S.H. (2016) Looking on the bright side of bias: Validation of an affective bias test for laboratory mice. Applied Animal Behaviour Science 181: 173–181	non-standard judgement bias test

22	Henry, S., Fureix, C., Rowberry, R., Bateson, M. (2017) Do horses with poor welfare show 'pessimistic' cognitive biases? The Science of Nature 104: 8	no affect manipulation or drug-based manipulation
23	Hernandez, C.E., Hinch, G., Lea, J., Ferguson, D., Lee, C. (2015) Acute stress enhances sensitivity to a highly attractive food reward without affecting judgement bias in laying hens. Applied Animal Behaviour Science 163: 135–143	data missing or not extractable
24	Hintze, S., Roth, E., Bachmann, I., Würbel, H. (2017) Toward a Choice-Based Judgment Bias Task for Horses. Journal of Applied Animal Welfare Science 20: 123–136	no affect manipulation or drug-based manipulation
25	Hymel, K.A., Sufka, K.J. (2012) Pharmacological reversal of cognitive bias in the chick anxiety-depression model. Neuropharmacology 62: 161–166	non-standard judgement bias test
26	Jones, S., Paul, E.S., Dayan, P., Robinson, E., Mendl, M. (2017). Pavlovian influences on learning differ between rats and mice in a counter-balanced Go/NoGo judgement bias task. Behavioural Brain Research 331: 214–224	no affect manipulation or drug-based manipulation
27	Karagiannis, C.I., Burman, O.H.P., Mills, D.S. (2015) Dogs with separation-related problems show a "less pessimistic" cognitive bias during treatment with fluoxetine (Reconcile TM) and a behaviour modification plan. BMC Veterinary Research 11: 42277	data missing or not extractable
28	Kis, A., Hernádi, A., Kanizsár, O., Gácsi, M., Topál, J. (2015) Oxytocin induces positive expectations about ambivalent stimuli (cognitive bias) in dogs. Hormones and Behavior 69: 42185	no affect manipulation or drug-based manipulation
29	Kregiel, J., Golebiowska, J., Popik, P., Rygula, R. (2016) Dopamine induces an optimism bias in rats-Pharmacological proof for the translational validity of the ambiguous-cue interpretation test. Behavioural Brain Research 297: 84–90	study retracted
30	Kregiel, J., Malek, N., Popik, P., Starowicz, K., Rygula, R. (2016) Anandamide mediates cognitive judgement bias in rats. Neuropharmacology 101: 146–153	no affect manipulation or drug-based manipulation
31	McGuire, M.C., Vonk, J. (2018) Gorillas (Gorilla gorilla gorilla) Fail to Learn Abstract Cues of Differential Outcomes in a Novel Cognitive Bias Test. Animal Behavior and Cognition 5: 103–117	non-standard judgement bias test
32	McGuire, M.C., Vonk, J., Fuller, G., Allard, S. (2017) Using an Ambiguous Cue Paradigm to Assess Cognitive Bias in Gorillas (*Gorilla gorilla gorilla*) during a Forage Manipulation. Animal Behavior and Cognition 4: 91–104	non-standard judgement bias test
33	McGuire, M.C., Vonk, J., Johnson-Ulrich, Z. (2017) Ambiguous Results When Using the Ambiguous-Cue Paradigm to Assess	no affect manipulation or drug-based manipulation

	Learning and Cognitive Bias in Gorillas and a Black Bear. Behavioral Sciences 7: 51	
34	McGuire, M.C., Williams, K.L., Welling, L L., Vonk, J. (2015). Cognitive bias in rats is not influenced by oxytocin. Frontiers in Psychology 6: 1306	no affect manipulation or drug-based manipulation
35	McHugh, S.B., Barkus, C., Lima, J., Glover, L.R., Sharp, T., Bannerman, D.M. (2015) SERT and uncertainty: serotonin transporter expression influences information processing biases for ambiguous aversive cues in mice. Genes Brain and Behavior 14: 330–336	non-standard judgement bias test
36	Mendl, M., Brooks, J., Basse, C., Burman, O., Paul, E., Blackwell, E., Casey, R. (2010) Dogs showing separation-related behaviour exhibit a 'pessimistic' cognitive bias. Current Biology 20: R839–R840	no affect manipulation or drug-based manipulation
37	Monk, J.E., Belson, S., Colditz, I.G., Lee, C. (2018) Attention Bias Test Differentiates Anxiety and Depression in Sheep. Frontiers in Behavioral Neuroscience 12: 246	no judgement bias test
38	Muehlemann, T., Reefmann, N., Wechsler, B., Wolf, M., Gygax, L. (2011) In vivo functional near-infrared spectroscopy measures mood-modulated cerebral responses to a positive emotional stimulus in sheep. NeuroImage 54: 1625–1633	no judgement bias test
39	Murphy, E., Kraak, L., van den Broek, J., Nordquist, R.E., van der Staay, F.J. (2014) Decision-making under risk and ambiguity in low-birth-weight pigs. Animal Cognition 18: 561–572	no affect manipulation or drug-based manipulation
40	Novak, J., Bailoo, J.D., Melotti, L., Rommen, J., Würbel, H. (2015) An exploration based cognitive bias test for mice: Effects of handling method and stereotypic behaviour. PLOS One 10: e0130718	data missing or not extractable
41	Novak, J., Bailoo, J.D., Melotti, L., Wurbel, H. (2016) Effect of Cage-Induced Stereotypies on Measures of Affective State and Recurrent Perseveration in CD-1 and C57BL/6 Mice. PLOS One 11: e0153203	no affect manipulation or drug-based manipulation
42	Pomerantz, O., Terkel, J., Suomi, S.J., Paukner, A. (2012) Stereotypic head twirls, but not pacing, are related to a 'pessimistic'-like judgment bias among captive tufted capuchins (*Cebus apella*). Animal Cognition 15: 689–698	no affect manipulation or drug-based manipulation
43	Rafa, D., Kregiel, J., Popik, P., Rygula, R. (2015) Effects of optimism on gambling in the rat slot machine task. Behavioural Brain Research 300: 97–105	no affect manipulation or drug-based manipulation
44	Rafa, D., Kregiel, J., Popik, P., Rygula, R. (2016) Effects of	no affect manipulation or

	optimism on gambling in the rat slot machine task. Behavioural Brain Research 300: 97–105	drug-based manipulation
45	Reefmann, N., Muehlemann, T., Wechsler, B., Gygax, L. (2012) Housing induced mood modulates reactions to emotional stimuli in sheep. Applied Animal Behaviour Science 136: 146–155	no judgement bias test
46	Roelofs, S., Boleij, H., Nordquist, R.E., van der Staay, F.J. (2016) Making decisions under ambiguity: Judgment bias tasks for assessing emotional state in animals. Frontiers in Behavioral Neuroscience 10: 119	no affect manipulation or drug-based manipulation
47	Roelofs, S., Nordquist, R.E., Staay, F.J. (2017) Female and male pigs' performance in a spatial holeboard and judgment bias task. Applied Animal Behaviour Science 191: 5–16	no affect manipulation or drug-based manipulation
48	Rygula, R., Golebiowska, J., Kregiel, J., Holuj, M., Popik, P. (2015) Acute administration of lithium, but not valproate, modulates cognitive judgment bias in rats. Psychopharmacology 232: 2149–2156	no affect manipulation or drug-based manipulation
49	Rygula, R., Golebiowska, J., Kregiel, J., Kubik, J., Popik, P. (2015) Effects of optimism on motivation in rats. Frontiers in Behavioral Neuroscience 9: 32	no affect manipulation or drug-based manipulation
50	Rygula, R., Popik, P. (2016) Trait "pessimism" is associated with increased sensitivity to negative feedback in rats. Cognitive Affective & Behavioral Neuroscience 16: 516–526	no affect manipulation or drug-based manipulation
51	Rygula, R., Szczech, E., Kregiel, J., Golebiowska, J., Kubik, J., Popik, P. (2015) Cognitive judgment bias in the psychostimulant-induced model of mania in rats. Psychopharmacology 232: 651–660	no affect manipulation or drug-based manipulation
52	Sahin, C., Doostdar, N., Neill, J.C. (2016) Towards the development of improved tests for negative symptoms of schizophrenia in a validated animal model. Behavioural Brain Research 312: 93–101	no affect manipulation or drug-based manipulation
53	Saito, Y., Yuki, S., Seki, Y., Kagawa, H., Okanoya, K. (2016) Cognitive bias in rats evoked by ultrasonic vocalizations suggests emotional contagion. Behavioural Processes 132: 42312	no affect manipulation or drug-based manipulation
54	Salmeto, A.L., Hymel, K. A., Carpenter, E.C., Brilot, B.O., Bateson, M., Sufka, K.J. (2011) Cognitive bias in the chick anxiety-depression model. Brain Research 1373: 124–130	no affect manipulation or drug-based manipulation
55	Starling, M. J., Branson, N., Cody, D., Starling, T.R., McGreevy, P.D. (2014). Canine sense and sensibility: tipping points and response latency variability as an optimism index in a canine judgement bias assessment. PLOS One 9: e107794	no affect manipulation or drug-based manipulation
56	Stuart, S.A., Butler, P., MunafÃ², M.R., Nutt, D.J., Robinson, E.S.J.	no judgement bias test

	(2013) A translational rodent assay of affective biases in depression and antidepressant therapy. Neuropsychopharmacology 38: 1625–1635	
57	Sümegi, Z., Gácsi, M., Topál, J. (2014) Conditioned placebo effect in dogs decreases separation related behaviours. Applied Animal Behaviour Science 159: 90–98	no affect manipulation or drug-based manipulation
58	Svendsen, P.M., Malmkvist, J., Halekoh, U., Mendl, M. (2012) Responses of mink to auditory stimuli: Prerequisites for applying the 'cognitive bias' approach. Behavioural Processes 91: 291–297	no judgement bias test
59	Tami, G., Torre, C., Compagnucci, M., Manteca, X. (2011) Interpretation of ambiguous spatial stimuli in cats. Animal Welfare 20: 185–189	no affect manipulation or drug-based manipulation
60	Titulaer, M., Blackwell, E.J., Mendl, M., Casey, R.A. (2013) Cross sectional study comparing behavioural, cognitive and physiological indicators of welfare between short and long term kennelled domestic dogs. Applied Animal Behaviour Science 147: 149–158	data missing or not extractable
61	Verbeek, E., Ferguson, D., Quinquet de Monjour, P., Lee, C. (2014) Generating positive affective states in sheep: The influence of food rewards and opioid administration. Applied Animal Behaviour Science 154: 39–47	no affect manipulation or drug-based manipulation
62	Vonk, J., McGuire, M.C., Johnson-Ulrich, Z. (2019) Seasonal changes in affect in an American black bear (*Ursus americanus*). Wildlife Biology 10: 277–284	no affect manipulation or drug-based manipulation

Table S4

Phylogenetic Meta-Analytical (intercept-only) model estimating the overall effect of the experimental manipulations on the judgement bias in animals. Mean = overall meta-analytic effect; CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). I^2 = estimates of heterogeneity values accounted for by different random effects. ArticleID = identity of an original paper from which data was extracted; ExperimentID = identity of an experiment from which data originated; Phylogeny = phylogenetic variance-covariance matrix representing evolutionary relationships among species; EffectID = residual variance. *N levels* = number of levels of each random effect.

Data		Fixed effects				Random effects	
		Mean	CI.lb	CI.ub		I^2	N levels
All data	Intercept	**0.201**	**0.028**	**0.374**	Total	76.4	459
					Phylogeny	2.0	22
					ScalePoint	0.4	5
					ArticleID	2.8	71
					ExperimentID	3.1	91
					EffectID	68.1	459
Ambiguous cues subset	Intercept	0.291	-0.002	0.584	Total	77.5	269
					Phylogeny	7.5	22
					ScalePoint	0	3
					ArticleID	9.3	71
					ExperimentID	14.6	91
					EffectID	46.1	269
Mid-cue effect subset	Intercept	0.291	-0.051	0.634	Total	78.4	108
					Phylogeny	9.4	22
					ScalePoint	0.0	1
					ArticleID	27.9	71
					ExperimentID	11.4	91
					EffectID	29.7	108
Largest absolute effect subset	Intercept	0.376	-0.049	0.801	Total	89.8	108
					Phylogeny	4.3	22
					ScalePoint	0.6	5
					ArticleID	0.0	71
					ExperimentID	47.5	91
					EffectID	37.4	108
Dominant effect subset	Intercept	**0.609**	**0.119**	**1.099**	Total	88.6	108
					Phylogeny	9.1	22
					ScalePoint	0.0	5

					ArticleID	0.0	71
					ExperimentID	19.8	91
					EffectID	59.6	108

Table S5

Meta-analytical (intercept-only) model estimating the overall effect of the experimental manipulations on the judgement bias in animals without controlling for phylogeny. Mean = overall meta-analytic effect; CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). SpeciesID = identity of a species from which data originated; ArticleID = identity of an original paper from which data was extracted; ExperimentID = identity of an experiment from which data originated; EffectID = residual variance. I^2 = estimates of heterogeneity values accounted for by different random effects. *N levels* = number of levels of each random effect.

Data		Fixed effects			Random effects		
		Mean	CI.lb	CI.ub		I^2	N levels
All data	Intercept	**0.204**	**0.087**	**0.320**	Total	76.3	459
					SpeciesID	0.0	22
					ScalePoint	0.3	5
					ArticleID	4.8	71
					ExperimentID	2.6	91
						68.5	459
Ambiguous cues subset	Intercept	**0.261**	**0.103**	**0.419**	Total	76.9	269
					SpeciesID	1.8	22
					ScalePoint	0.0	3
					ArticleID	13.3	71
					ExperimentID	14.3	91
					EffectID	47.4	269
Mid-cue effect subset	Intercept	**0.251**	**0.056**	**0.446**	Total	77.6	108
					SpeciesID	2.8	22
					ScalePoint	0.0	1
					ArticleID	33.3	71
					ExperimentID	9.9	91
					EffectID	31.6	108
Largest absolute effect subset	Intercept	**0.386**	**0.080**	**0.692**	Total	89.7	108
					SpeciesID	1.1	22
					ScalePoint	1.1	5
					ArticleID	6.7	71
					ExperimentID	42.9	91
					EffectID	38.1	108

Dominant effect subset	Intercept	**0.577**	**0.326**	**0.827**	Total	88.2	108
					SpeciesID	0.0	22
					ScalePoint	0.0	5
					ArticleID	8.1	71
					ExperimentID	15.5	91
					EffectID	64.5	108

Table S6

Univariate Multilevel Phylogenetic Meta-Regression models with identity of species used in judgement bias tests as a fixed effect. Intercept values are shown for each species. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. Mean = overall meta-analytic effect for each species (intercept); CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Species name	Fixed effects			R^2
		Mean	CI.lb	CI.ub	
All data					0.070
	Apis mellifera carnica	0.237	-0.698	1.173	15
	Bombus terrestris audax	0.538	-0.732	1.808	5
	Bos taurus	0.335	-0.646	1.317	15
	Callithrix jacchus	0.569	-0.889	2.026	3
	Canis lupus familiaris	0.198	-0.654	1.050	17
	Capra hircus	-0.193	-1.229	0.842	15
	Coturnix japonica	-0.333	-1.362	0.696	20
	Drosophila melanogaster	0.651	-1.132	2.434	3
	Equus caballus	-0.217	-1.240	0.806	14
	Felis catus	0.653	-0.745	2.052	5
	Gallus gallus	-0.033	-0.858	0.792	41
	Macaca mulatta	0.433	-0.745	1.612	10
	Mesocricetus auratus	0.324	-0.807	1.455	5
	Mus musculus	0.271	-0.640	1.181	43
	Ovis aries	0.163	-0.606	0.932	65
	Pecari tajacu	0.069	-1.289	1.427	3
	Rattus norvegicus	0.401	-0.357	1.159	76
	Serinus canaria	0.263	-0.881	1.407	5
	Sturnus vulgaris	0.066	-0.761	0.894	40
	Sus scrofa domesticus	0.223	-0.565	1.010	51
	Tayassu pecari	0.935	-0.436	2.307	3
	Ursus arctos horribilis	0.403	-0.830	1.636	5
Ambiguous cues subset					0.186
	Apis mellifera carnica	0.346	-0.642	1.334	9

	Bombus terrestris audax	1.052	-0.380	2.483	3
	Bos taurus	0.594	-0.443	1.632	9
	Callithrix jacchus	0.829	-1.037	2.695	1
	Canis lupus familiaris	0.298	-0.583	1.179	11
	Capra hircus	-0.433	-1.557	0.691	9
	Coturnix japonica	-0.724	-1.776	0.328	12
	Drosophila melanogaster	**2.457**	**0.146**	**4.767**	1
	Equus caballus	-0.400	-1.496	0.697	8
	Felis catus	0.683	-0.869	2.236	3
	Gallus gallus	-0.061	-0.897	0.775	23
	Macaca mulatta	0.681	-0.648	2.01	6
	Mesocricetus auratus	0.513	-0.774	1.801	3
	Mus musculus	0.240	-0.729	1.208	25
	Ovis aries	0.110	-0.663	0.882	39
	Pecari tajacu	-0.230	-1.966	1.506	1
	Rattus norvegicus	0.544	-0.211	1.299	44
	Serinus canaria	0.410	-0.890	1.709	3
	Sturnus vulgaris	0.182	-0.654	1.019	24
	Sus scrofa domesticus	0.386	-0.417	1.189	31
	Tayassu pecari	1.022	-0.744	2.788	1
	Ursus arctos horribilis	0.307	-1.075	1.688	3
Mid-cue effect subset					0.270
	Apis mellifera carnica	0.327	-0.804	1.459	3
	Bombus terrestris audax	1.569	-0.209	3.347	1
	Bos taurus	0.797	-0.396	1.991	3
	Callithrix jacchus	0.829	-1.025	2.683	1
	Canis lupus familiaris	0.334	-0.558	1.226	5
	Capra hircus	-0.396	-1.684	0.892	3
	Coturnix japonica	-0.845	-2.205	0.515	4
	Drosophila melanogaster	**2.457**	**0.156**	**4.757**	1
	Equus caballus	-0.673	-1.906	0.559	3
	Felis catus	0.443	-1.436	2.323	1
	Gallus gallus	-0.166	-1.048	0.716	9
	Macaca mulatta	0.786	-0.780	2.352	2
	Mesocricetus auratus	-0.042	-1.602	1.518	1
	Mus musculus	0.276	-0.733	1.285	9
	Ovis aries	0.030	-0.710	0.771	13
	Pecari tajacu	-0.230	-1.952	1.492	1
	Rattus norvegicus	0.600	-0.087	1.287	24
	Serinus canaria	0.796	-0.801	2.392	1
	Sturnus vulgaris	0.081	-0.787	0.950	8
	Sus scrofa domesticus	0.261	-0.508	1.029	13
	Tayassu pecari	1.022	-0.731	2.775	1
	Ursus arctos horribilis	-0.079	-1.748	1.591	1
Largest absolute effect subset					0.218
	Apis mellifera carnica	1.353	-0.480	3.186	3
	Bombus terrestris audax	1.569	-1.237	4.375	1
	Bos taurus	**2.477**	**0.389**	**4.565**	3
	Callithrix jacchus	0.829	-2.025	3.683	1

	Canis lupus familiaris	0.385	-1.177	1.948	5
	Capra hircus	-0.395	-2.468	1.677	3
	Coturnix japonica	-0.902	-2.668	0.865	4
	Drosophila melanogaster	-3.069	-6.358	0.219	1
	Equus caballus	-1.076	-3.011	0.858	3
	Felis catus	1.195	-1.711	4.100	1
	Gallus gallus	-0.130	-1.566	1.306	9
	Macaca mulatta	0.878	-1.650	3.407	2
	Mesocricetus auratus	1.111	-1.577	3.799	1
	Mus musculus	0.415	-1.234	2.064	9
	Ovis aries	0.094	-1.229	1.417	13
	Pecari tajacu	0.314	-2.460	3.088	1
	Rattus norvegicus	0.714	-0.515	1.943	24
	Serinus canaria	0.796	-1.899	3.490	1
	Sturnus vulgaris	0.132	-1.366	1.629	8
	Sus scrofa domesticus	0.682	-0.684	2.049	13
	Tayassu pecari	1.022	-1.768	3.812	1
	Ursus arctos horribilis	0.740	-2.026	3.505	1
Dominant effect subset					0.190
	Apis mellifera carnica	1.354	-0.366	3.074	3
	Bombus terrestris audax	1.569	-1.086	4.224	1
	Bos taurus	**2.411**	**0.430**	**4.391**	3
	Callithrix jacchus	0.829	-1.877	3.535	1
	Canis lupus familiaris	0.384	-1.075	1.843	5
	Capra hircus	-0.392	-2.301	1.516	3
	Coturnix japonica	-0.901	-2.566	0.764	4
	Drosophila melanogaster	2.457	-0.573	5.486	1
	Equus caballus	-0.844	-2.672	0.984	3
	Felis catus	1.195	-1.565	3.954	1
	Gallus gallus	0.149	-1.176	1.474	9
	Macaca mulatta	0.879	-1.377	3.135	2
	Mesocricetus auratus	1.111	-1.419	3.641	1
	Mus musculus	0.874	-0.630	2.377	9
	Ovis aries	0.684	-0.540	1.909	13
	Pecari tajacu	0.314	-2.307	2.936	1
	Rattus norvegicus	0.744	-0.391	1.879	24
	Serinus canaria	0.796	-1.742	3.333	1
	Sturnus vulgaris	0.110	-1.293	1.513	8
	Sus scrofa domesticus	0.621	-0.634	1.876	13
	Tayassu pecari	1.022	-1.616	3.660	1
	Ursus arctos horribilis	0.740	-1.872	3.352	1

Table S7

Univariate Multilevel Phylogenetic Meta-Regression models with sex of animals used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels.

Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	K
All data					0.024
	Females intercept	0.104	-0.063	0.271	225
	Males intercept	**0.365**	**0.155**	**0.575**	118
	Mixed-sex intercept	0.202	-0.009	0.414	116
	Females – Males contrast	0.261	-0.001	0.522	
	Females – Mixed-sex contrast	0.098	-0.165	0.361	
	Males – Mixed-sex contrast	-0.163	-0.455	0.130	
Ambiguous cues subset					0.052
	Females intercept	0.124	-0.133	0.382	133
	Males intercept	**0.523**	**0.194**	**0.853**	68
	Mixed-sex intercept	0.294	-0.015	0.603	68
	Females – Males contrast	**0.399**	**0.032**	**0.766**	
	Females – Mixed-sex contrast	0.170	-0.180	0.519	
	Males – Mixed-sex contrast	-0.229	-0.640	0.181	
Mid-cue effect subset					0.094
	Females intercept	0.052	-0.217	0.320	49
	Males intercept	**0.574**	**0.256**	**0.892**	32
	Mixed-sex intercept	0.238	-0.093	0.569	27
	Females – Males contrast	**0.522**	**0.112**	**0.933**	
	Females – Mixed-sex contrast	0.186	-0.240	0.612	
	Males – Mixed-sex contrast	-0.336	-0.795	0.123	
Largest absolute effect subset					0.067
	Females intercept	0.153	-0.263	0.570	49
	Males intercept	**0.875**	**0.388**	**1.361**	32
	Mixed-sex intercept	0.215	-0.320	0.749	27
	Females – Males contrast	**0.721**	**0.114**	**1.329**	
	Females – Mixed-sex contrast	0.061	-0.596	0.719	
	Males – Mixed-sex contrast	-0.660	-1.361	0.041	
Dominant effect subset					0.025
	Females intercept	0.514	-0.009	1.038	49
	Males intercept	**0.902**	**0.248**	**1.557**	32
	Mixed-sex intercept	0.484	-0.133	1.102	27
	Females – Males contrast	0.388	-0.270	1.046	
	Females – Mixed-sex contrast	-0.030	-0.650	0.591	
	Males – Mixed-sex contrast	-0.418	-1.163	0.327	

Table S8

Univariate Multilevel Phylogenetic Meta-Regression models with type of task used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.021
	Active choice intercept	**0.432**	**0.151**	**0.712**	70
	Go/no-go intercept	**0.154**	**0.005**	**0.304**	389
	Active choice – Go/no-go contrast	-0.277	-0.567	0.012	
Ambiguous cues subset					0.059
	Active choice intercept	**0.688**	**0.297**	**1.080**	38
	Go/no-go intercept	0.193	-0.042	0.428	231
	Active choice – Go/no-go contrast	-0.495	-0.878	-0.113	
Mid-cue effect subset					0.095
	Active choice intercept	**0.708**	**0.286**	**1.130**	22
	Go/no-go intercept	0.160	-0.080	0.400	86
	Active choice – Go/no-go contrast	-0.548	-0.992	-0.103	
Largest absolute effect subset					0.008
	Active choice intercept	0.590	-0.071	1.251	22
	Go/no-go intercept	0.327	-0.076	0.731	86
	Active choice – Go/no-go contrast	-0.263	-0.939	0.414	
Dominant effect subset					0.015
	Active choice intercept	**0.882**	**0.184**	**1.579**	22
	Go/no-go intercept	**0.539**	**0.065**	**1.014**	86
	Active choice – Go/no-go contrast	-0.342	-0.985	0.301	

Table S9

Univariate Multilevel Phylogenetic Meta-Regression models with type of the cue used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	*k*
All data					0.044
	Auditory cue intercept	**0.393**	**0.136**	**0.651**	81
	Olfactory cue intercept	0.280	-0.215	0.775	18
	Spatial cue intercept	0.165	-0.003	0.334	177
	Tactile cue intercept	**0.658**	**0.136**	**1.180**	16
	Visual cue intercept	0.067	-0.133	0.268	167
	Auditory – Olfactory contrast	-0.113	-0.668	0.441	
	Auditory – Spatial contrast	-0.228	-0.530	0.074	
	Auditory – Tactile contrast	0.265	-0.313	0.843	
	Auditory – Visual contrast	**-0.326**	**-0.647**	**-0.005**	
	Olfactory – Spatial contrast	-0.115	-0.634	0.405	
	Olfactory – Tactile contrast	0.378	-0.338	1.095	
	Olfactory – Visual contrast	-0.213	-0.743	0.318	
	Spatial – Tactile contrast	0.493	-0.052	1.038	
	Spatial – Visual contrast	-0.098	-0.353	0.157	
	Tactile – Visual contrast	**-0.591**	**-1.147**	**-0.035**	
Ambiguous cues subset					0.082
	Auditory cue intercept	**0.626**	**0.261**	**0.99**	39
	Olfactory cue intercept	0.551	-0.130	1.232	10
	Spatial cue intercept	0.148	-0.066	0.361	111
	Tactile cue intercept	**0.693**	**0.106**	**1.280**	12
	Visual cue intercept	0.125	-0.139	0.388	97
	Auditory – Olfactory contrast	-0.075	-0.847	0.698	
	Auditory – Spatial contrast	**-0.478**	**-0.900**	**-0.056**	
	Auditory – Tactile contrast	0.067	-0.624	0.758	
	Auditory – Visual contrast	**-0.501**	**-0.951**	**-0.051**	
	Olfactory – Spatial contrast	-0.403	-1.117	0.310	
	Olfactory – Tactile contrast	0.142	-0.757	1.041	
	Olfactory – Visual contrast	-0.426	-1.157	0.304	
	Spatial – Tactile contrast	0.545	-0.080	1.170	
	Spatial – Visual contrast	-0.023	-0.362	0.316	
	Tactile – Visual contrast	-0.568	-1.212	0.075	
Mid-cue effect subset					0.133
	Auditory cue intercept	**0.625**	**0.209**	**1.041**	21
	Olfactory cue intercept	0.666	-0.209	1.540	4
	Spatial cue intercept	0.060	-0.209	0.330	40
	Tactile cue intercept	**0.808**	**0.147**	**1.469**	8
	Visual cue intercept	0.145	-0.191	0.481	35
	Auditory – Olfactory contrast	0.041	-0.928	1.009	
	Auditory – Spatial contrast	**-0.565**	**-1.061**	**-0.069**	
	Auditory – Tactile contrast	0.183	-0.598	0.964	
	Auditory – Visual contrast	-0.480	-1.015	0.055	
	Olfactory – Spatial contrast	-0.605	-1.520	0.310	
	Olfactory – Tactile contrast	0.142	-0.954	1.238	
	Olfactory – Visual contrast	-0.521	-1.457	0.416	
	Spatial – Tactile contrast	**0.747**	**0.033**	**1.461**	
	Spatial – Visual contrast	0.085	-0.346	0.515	

	Tactile – Visual contrast	-0.663	-1.404	0.079	
Largest absolute effect subset					0.061
	Auditory cue intercept	**0.816**	**0.200**	**1.432**	21
	Olfactory cue intercept	0.561	-0.754	1.875	4
	Spatial cue intercept	0.166	-0.258	0.591	40
	Tactile cue intercept	**1.037**	**0.076**	**1.997**	8
	Visual cue intercept	0.243	-0.273	0.759	35
	Auditory – Olfactory contrast	-0.255	-1.702	1.192	
	Auditory – Spatial contrast	-0.650	-1.388	0.088	
	Auditory – Tactile contrast	0.221	-0.911	1.352	
	Auditory – Visual contrast	-0.573	-1.368	0.222	
	Olfactory – Spatial contrast	-0.395	-1.772	0.983	
	Olfactory – Tactile contrast	0.476	-1.150	2.102	
	Olfactory – Visual contrast	-0.318	-1.724	1.088	
	Spatial – Tactile contrast	0.870	-0.170	1.911	
	Spatial – Visual contrast	0.077	-0.581	0.734	
	Tactile – Visual contrast	-0.794	-1.878	0.291	
Dominant effect subset					0.077
	Auditory cue intercept	**0.809**	**0.237**	**1.380**	21
	Olfactory cue intercept	**1.579**	**0.353**	**2.805**	4
	Spatial cue intercept	0.378	-0.025	0.780	40
	Tactile cue intercept	**1.154**	**0.268**	**2.040**	8
	Visual cue intercept	0.386	-0.090	0.862	35
	Auditory – Olfactory contrast	0.770	-0.582	2.123	
	Auditory – Spatial contrast	-0.431	-1.104	0.242	
	Auditory – Tactile contrast	0.346	-0.684	1.375	
	Auditory – Visual contrast	-0.422	-1.148	0.303	
	Olfactory – Spatial contrast	-1.201	-2.490	0.088	
	Olfactory – Tactile contrast	-0.425	-1.937	1.088	
	Olfactory – Visual contrast	-1.193	-2.508	0.122	
	Spatial – Tactile contrast	0.776	-0.181	1.734	
	Spatial – Visual contrast	0.009	-0.592	0.609	
	Tactile – Visual contrast	-0.768	-1.760	0.224	

Table S10

Univariate Multilevel Phylogenetic Meta-Regression models with cue reinforcement scheme used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.030
	Reward-Null intercept	0.052	-0.192	0.295	132
	Reward-Punishment intercept	**0.216**	**0.036**	**0.396**	283
	Reward-Reward intercept	**0.488**	**0.137**	**0.839**	44
	Reward-Null – Reward-Punishment contrast	0.164	-0.087	0.415	
	Reward-Null – Reward-Reward contrast	**0.436**	**0.045**	**0.827**	
	Reward-Punishment – Reward-Reward contrast	0.272	-0.082	0.626	
Ambiguous cues subset					0.062
	Reward-Null intercept	0.046	-0.313	0.405	82
	Reward-Punishment intercept	**0.325**	**0.030**	**0.620**	159
	Reward-Reward intercept	**0.654**	**0.187**	**1.121**	28
	Reward-Null – Reward-Punishment contrast	0.279	-0.052	0.609	
	Reward-Null – Reward-Reward contrast	**0.608**	**0.122**	**1.094**	
	Reward-Punishment – Reward-Reward contrast	0.329	-0.110	0.768	
Mid-cue effect subset					0.075
	Reward-Null intercept	-0.024	-0.461	0.413	29
	Reward-Punishment intercept	**0.352**	**0.012**	**0.691**	65
	Reward-Reward intercept	**0.595**	**0.046**	**1.144**	14
	Reward-Null – Reward-Punishment contrast	0.376	-0.050	0.801	
	Reward-Null – Reward-Reward contrast	**0.619**	**0.015**	**1.224**	
	Reward-Punishment – Reward-Reward contrast	0.243	-0.289	0.776	
Largest absolute effect subset					0.045
	Reward-Null intercept	-0.053	-0.640	0.534	29
	Reward-Punishment intercept	**0.519**	**0.110**	**0.929**	65
	Reward-Reward intercept	0.580	-0.199	1.359	14
	Reward-Null – Reward-Punishment contrast	0.572	-0.070	1.215	
	Reward-Null – Reward-Reward contrast	0.633	-0.288	1.554	
	Reward-Punishment – Reward-Reward contrast	0.061	-0.757	0.879	
Dominant effect subset					0.059
	Reward-Null intercept	0.139	-0.461	0.738	29
	Reward-Punishment intercept	**0.780**	**0.306**	**1.253**	65
	Reward-Reward intercept	0.710	-0.070	1.491	14
	Reward-Null – Reward-Punishment contrast	**0.641**	**0.061**	**1.222**	
	Reward-Null – Reward-Reward contrast	0.572	-0.274	1.417	
	Reward-Punishment – Reward-Reward contrast	-0.069	-0.821	0.682	

Table S11

Univariate Multilevel Phylogenetic Meta-Regression models with reinforcement category for ambiguous cues used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.001
	No intercept	**0.204**	**0.026**	**0.382**	421
	Yes intercept	0.124	-0.490	0.738	38
	No – Yes contrast	0.080	-0.527	0.686	
Ambiguous cues subset					0.003
	No intercept	0.297	-0.004	0.598	247
	Yes intercept	0.156	-0.666	0.978	22
	No – Yes contrast	0.141	-0.650	0.931	
Mid-cue effect subset					0.014
	No intercept	0.306	-0.054	0.666	100
	Yes intercept	-0.031	-0.993	0.932	8
	No – Yes contrast	0.337	-0.587	1.261	
Largest absolute effect subset					0.069
	No intercept	0.421	-0.067	0.909	100
	Yes intercept	-0.832	-2.248	0.585	8
	No – Yes contrast	1.253	-0.122	2.628	
Dominant effect subset					0.02
	No intercept	**0.636**	**0.119**	**1.153**	100
	Yes intercept	0.015	-1.275	1.305	8
	No – Yes contrast	0.622	-0.604	1.847	

Table S12

Univariate Multilevel Phylogenetic Meta-Regression models with position of the cue (i.e. ScalePoint) used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered

statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.014
	P cue intercept	0.063	-0.153	0.278	95
	NP cue intercept	0.225	-0.002	0.451	81
	MID cue intercept	**0.250**	**0.042**	**0.458**	108
	NN cue intercept	**0.303**	**0.075**	**0.530**	80
	N cue intercept	0.181	-0.034	0.396	95
	P – NP contrast	0.162	-0.057	0.381	
	P – MID contrast	0.187	-0.014	0.389	
	P – NN contrast	**0.240**	**0.020**	**0.460**	
	P – N contrast	0.118	-0.088	0.325	
	NP – MID contrast	0.025	-0.188	0.238	
	NP – NN contrast	0.078	-0.148	0.304	
	NP – N contrast	-0.044	-0.262	0.175	
	MID – NN contrast	-0.025	-0.238	0.188	
	MID – N contrast	0.053	-0.161	0.267	
	NN – N contrast	-0.122	-0.341	0.098	
Ambiguous cues subset					0.003
	P cue intercept				0
	NP cue intercept	0.284	-0.042	0.610	81
	MID cue intercept	0.260	-0.054	0.575	108
	NN cue intercept	**0.356**	**0.030**	**0.683**	80
	N cue intercept				0
	P – NP contrast				
	P – MID contrast				
	P – NN contrast				
	P – N contrast				
	NP – MID contrast	-0.024	-0.217	0.170	
	NP – NN contrast	0.072	-0.129	0.274	
	NP – N contrast				
	MID – NN contrast	0.096	-0.098	0.290	
	MID – N contrast				
	NN – N contrast				
Mid-cue effect subset					0
	P cue intercept				0
	NP cue intercept				0
	MID cue intercept	0.291	-0.051	0.634	108
	NN cue intercept				0
	N cue intercept				0
	P – NP contrast				
	P – MID contrast				
	P – NN contrast				
	P – N contrast				
	NP – MID contrast				

	NP – NN contrast				
	NP – N contrast				
	MID – NN contrast				
	MID – N contrast				
	NN – N contrast				
Largest absolute effect subset					0.052
	P cue intercept	-0.054	-0.664	0.557	26
	NP cue intercept	0.517	-0.209	1.244	13
	MID cue intercept	**0.708**	**0.205**	**1.212**	31
	NN cue intercept	0.410	-0.196	1.016	22
	N cue intercept	0.270	-0.380	0.920	16
	P – NP contrast	0.571	-0.330	1.472	
	P – MID contrast	**0.762**	**0.035**	**1.490**	
	P – NN contrast	0.464	-0.350	1.278	
	P – N contrast	0.324	-0.528	1.176	
	NP – MID contrast	0.191	-0.597	0.979	
	NP – NN contrast	-0.107	-1.011	0.796	
	NP – N contrast	-0.247	-1.173	0.679	
	MID – NN contrast	-0.191	-0.979	0.597	
	MID – N contrast	-0.298	-1.041	0.444	
	NN – N contrast	-0.140	-0.974	0.695	
Dominant effect subset					0.033
	P cue intercept	0.161	-0.550	0.871	19
	NP cue intercept	0.752	-0.068	1.571	12
	MID cue intercept	**0.733**	**0.140**	**1.325**	34
	NN cue intercept	**0.698**	**0.037**	**1.358**	28
	N cue intercept	0.647	-0.073	1.368	15
	P – NP contrast	0.591	-0.305	1.487	
	P – MID contrast	0.572	-0.132	1.276	
	P – NN contrast	0.537	-0.228	1.303	
	P – N contrast	0.487	-0.341	1.315	
	NP – MID contrast	-0.019	-0.804	0.766	
	NP – NN contrast	-0.054	-0.913	0.806	
	NP – N contrast	-0.104	-1.017	0.808	
	MID – NN contrast	0.019	-0.766	0.804	
	MID – N contrast	-0.035	-0.697	0.627	
	NN – N contrast	-0.050	-0.828	0.727	

Table S13

Univariate Multilevel Phylogenetic Meta-Regression models with source of animals used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates

effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.004
	Captive intercept	**0.224**	**0.040**	**0.408**	
	Wild-caught intercept	0.073	-0.295	0.441	
	Captives – Wild-caught contrast	-0.151	-0.535	0.233	
Ambiguous cues subset					0.000
	Captive intercept	0.302	-0.023	0.627	
	Wild-caught intercept	0.261	-0.270	0.791	
	Captives – Wild-caught contrast	-0.042	-0.579	0.496	
Mid-cue effect subset					0.003
	Captive intercept	0.323	-0.065	0.711	
	Wild-caught intercept	0.180	-0.492	0.852	
	Captives – Wild-caught contrast	-0.143	-0.831	0.546	
Largest absolute effect subset					0.006
	Captive intercept	0.323	-0.173	0.819	
	Wild-caught intercept	0.682	-0.328	1.691	
	Captives – Wild-caught contrast	0.358	-0.682	1.399	
Dominant effect subset					0.000
	Captive intercept	**0.615**	**0.079**	**1.151**	
	Wild-caught intercept	0.610	-0.369	1.588	
	Captives – Wild-caught contrast	-0.005	-1.001	0.991	

Table S14

Univariate Multilevel Phylogenetic Meta-Regression models with age of animals used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.000
	Adults intercept	**0.210**	**0.025**	**0.395**	
	Juveniles intercept	0.179	-0.068	0.426	
	Adults – Juveniles contrast	-0.032	-0.279	0.215	
Ambiguous cues subset					0.001
	Adults intercept	0.285	-0.036	0.606	

	Juveniles intercept	0.325	-0.078	0.728	
	Adults – Juveniles contrast	0.040	-0.301	0.382	
Mid-cue effect subset					0.000
	Adults intercept	0.290	-0.082	0.663	
	Juveniles intercept	0.308	-0.177	0.793	
	Adults – Juveniles contrast	0.018	-0.414	0.449	
Largest absolute effect subset					0.013
	Adults intercept	0.298	-0.236	0.832	
	Juveniles intercept	0.615	-0.088	1.317	
	Adults – Juveniles contrast	0.316	-0.327	0.960	
Dominant effect subset					0.003
	Adults intercept	**0.586**	**0.033**	**1.139**	
	Juveniles intercept	**0.719**	**0.022**	**1.417**	
	Adults – Juveniles contrast	0.133	-0.465	0.732	

Table S15

Univariate Multilevel Phylogenetic Meta-Regression models with type of affect manipulation used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.008
	Enrichment intercept	0.112	-0.129	0.353	157
	Stress intercept	**0.244**	**0.044**	**0.444**	302
	Enrichment – Stress contrast	0.132	-0.097	0.361	
Ambiguous cues subset					0.008
	Enrichment intercept	0.201	-0.174	0.576	93
	Stress intercept	**0.339**	**0.011**	**0.667**	176
	Enrichment – Stress contrast	0.138	-0.172	0.448	
Mid-cue effect subset					0.02
	Enrichment intercept	0.150	-0.296	0.596	37
	Stress intercept	0.367	-0.020	0.753	71
	Enrichment – Stress contrast	0.217	-0.162	0.595	
Largest absolute effect subset					0.008
	Enrichment intercept	0.219	-0.370	0.808	37
	Stress intercept	0.449	-0.027	0.924	71
	Enrichment – Stress contrast	0.230	-0.345	0.804	

Dominant effect subset					0.021
	Enrichment intercept	0.372	-0.234	0.978	37
	Stress intercept	**0.721**	**0.203**	**1.240**	71
	Enrichment – Stress contrast	0.349	-0.171	0.870	

Table S16

Univariate Multilevel Phylogenetic Meta-Regression models with timing of affect manipulation used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.002
	Before/during intercept	**0.238**	**0.029**	**0.446**	167
	Long-term intercept	0.173	-0.024	0.369	292
	Before/during – Long-term contrast	-0.065	-0.280	0.150	
Ambiguous cues subset					0.003
	Before/during intercept	**0.334**	**0.004**	**0.663**	93
	Long-term intercept	0.255	-0.063	0.572	176
	Before/during – Long-term contrast	-0.079	-0.374	0.215	
Mid-cue effect subset					0.008
	Before/during intercept	0.362	-0.026	0.751	39
	Long-term intercept	0.230	-0.142	0.603	69
	Before/during – Long-term contrast	-0.132	-0.490	0.226	
Largest absolute effect subset					0
	Before/during intercept	0.404	-0.123	0.930	39
	Long-term intercept	0.357	-0.128	0.843	69
	Before/during – Long-term contrast	-0.046	-0.600	0.507	
Dominant effect subset					0.016
	Before/during intercept	**0.771**	**0.240**	**1.301**	39
	Long-term intercept	0.468	-0.033	0.969	69
	Before/during – Long-term contrast	-0.303	-0.813	0.207	

Table S17

Univariate Multilevel Phylogenetic Meta-Regression models with type of comparison of affect manipulations used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.010
	Benign-Worse intercept	**0.263**	**0.062**	**0.464**	230
	Better-Benign intercept	0.149	-0.104	0.402	135
	Better-Worse intercept	0.102	-0.162	0.365	94
	Benign-Worse – Better-Benign contrast	-0.114	-0.361	0.133	
	Benign-Worse – Better-Worse contrast	-0.162	-0.414	0.090	
	Better-Benign – Better-Worse contrast	-0.048	-0.347	0.251	
Ambiguous cues subset					0.007
	Benign-Worse intercept	**0.338**	**0.017**	**0.659**	132
	Better-Benign intercept	0.275	-0.100	0.650	81
	Better-Worse intercept	0.183	-0.218	0.583	56
	Benign-Worse – Better-Benign contrast	-0.063	-0.381	0.256	
	Benign-Worse – Better-Worse contrast	-0.155	-0.493	0.183	
	Better-Benign – Better-Worse contrast	-0.092	-0.488	0.303	
Mid-cue effect subset					0.007
	Benign-Worse intercept	0.329	-0.049	0.706	55
	Better-Benign intercept	0.309	-0.145	0.763	31
	Better-Worse intercept	0.166	-0.317	0.649	22
	Benign-Worse – Better-Benign contrast	-0.019	-0.431	0.392	
	Benign-Worse – Better-Worse contrast	-0.163	-0.590	0.265	
	Better-Benign – Better-Worse contrast	-0.144	-0.652	0.364	
Largest absolute effect subset					0.006
	Benign-Worse intercept	0.433	-0.032	0.898	55
	Better-Benign intercept	0.221	-0.383	0.824	31
	Better-Worse intercept	0.414	-0.246	1.073	22
	Benign-Worse – Better-Benign contrast	-0.212	-0.830	0.406	
	Benign-Worse – Better-Worse contrast	-0.019	-0.684	0.646	

	Better-Benign – Better-Worse contrast	0.193	-0.577	0.963	
Dominant effect subset					0.012
	Benign-Worse intercept	**0.713**	**0.194**	**1.232**	55
	Better-Benign intercept	0.484	-0.144	1.113	31
	Better-Worse intercept	0.444	-0.236	1.123	22
	Benign-Worse – Better-Benign contrast	-0.229	-0.807	0.349	
	Benign-Worse – Better-Worse contrast	-0.269	-0.882	0.343	
	Better-Benign – Better-Worse contrast	-0.040	-0.758	0.677	

Table S18

Univariate Multilevel Phylogenetic Meta-Regression models with experimental design type used in judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.008
	Between intercept	0.151	-0.061	0.364	302
	Within intercept	**0.279**	**0.044**	**0.514**	157
	Between – Within contrast	-0.128	-0.353	0.097	
Ambiguous cues subset					0.03
	Between intercept	0.199	-0.16	0.559	182
	Within intercept	**0.470**	**0.079**	**0.861**	87
	Between – Within contrast	-0.271	-0.578	0.037	
Mid-cue effect subset					0.024
	Between intercept	0.210	-0.205	0.625	71
	Within intercept	0.453	-0.001	0.908	37
	Between – Within contrast	-0.243	-0.618	0.132	
Largest absolute effect subset					0.049
	Between intercept	0.168	-0.420	0.755	71
	Within intercept	**0.756**	**0.098**	**1.415**	37
	Between – Within contrast	**-0.589**	**-1.158**	**-0.02**	
Dominant effect subset					0.030
	Between intercept	0.487	-0.170	1.143	71
	Within intercept	**0.921**	**0.206**	**1.636**	37
	Between – Within contrast	-0.434	-0.967	0.098	

Table S19

Univariate Multilevel Phylogenetic Meta-Regression models with food deprivation during judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	K
All data					0.000
	No intercept	**0.201**	**0.011**	**0.391**	324
	Yes intercept	0.202	-0.039	0.444	135
	No – Yes contrast	-0.001	-0.243	0.240	
Ambiguous cues subset					0.003
	No intercept	0.262	-0.038	0.561	192
	Yes intercept	0.347	-0.010	0.704	77
	No – Yes contrast	-0.085	-0.407	0.236	
Mid-cue effect subset					0.022
	No intercept	0.210	-0.119	0.540	77
	Yes intercept	**0.448**	**0.037**	**0.858**	31
	No – Yes contrast	-0.237	-0.628	0.154	
Largest absolute effect subset					0.083
	No intercept	0.167	-0.169	0.503	77
	Yes intercept	**0.948**	**0.439**	**1.458**	31
	No – Yes contrast	-0.781	-1.360	-0.202	
Dominant effect subset					0.033
	No intercept	**0.454**	**0.039**	**0.869**	77
	Yes intercept	**0.909**	**0.361**	**1.457**	31
	No – Yes contrast	-0.455	-1.002	0.092	

Table S20

Univariate Multilevel Phylogenetic Meta-Regression models with measurement type used during judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.000
	Latency intercept	**0.193**	**0.001**	**0.386**	201
	Proportion intercept	**0.192**	**0.006**	**0.378**	258
	Latency – Proportion contrast	0.002	-0.201	0.205	
Ambiguous cues subset					0.001
	Latency intercept	0.266	-0.054	0.585	123
	Proportion intercept	0.311	-0.001	0.624	146
	Latency – Proportion contrast	0.046	-0.221	0.312	
Mid-cue effect subset					0.009
	Latency intercept	0.211	-0.152	0.575	46
	Proportion intercept	0.348	-0.003	0.700	62
	Latency – Proportion contrast	0.137	-0.198	0.471	
Largest absolute effect subset					0.008
	Latency intercept	0.264	-0.189	0.717	46
	Proportion intercept	**0.484**	**0.051**	**0.917**	62
	Latency – Proportion contrast	0.220	-0.285	0.726	
Dominant effect subset					0.032
	Latency intercept	0.369	-0.083	0.820	46
	Proportion intercept	**0.777**	**0.342**	**1.212**	62
	Latency – Proportion contrast	0.409	-0.076	0.893	

Table S21

Univariate Multilevel Phylogenetic Meta-Regression models with automation of measurements during judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.000
	No intercept	**0.196**	**0.016**	**0.375**	388
	Yes intercept	0.237	-0.079	0.554	71
	No – Yes contrast	-0.042	-0.353	0.270	
Ambiguous cues subset					0.000
	No intercept	0.289	-0.011	0.588	232
	Yes intercept	0.314	-0.165	0.792	37
	No – Yes contrast	-0.025	-0.459	0.409	

Mid-cue effect subset					0.002
	No intercept	0.280	-0.068	0.628	91
	Yes intercept	0.363	-0.210	0.937	17
	No – Yes contrast	-0.083	-0.614	0.448	
Largest absolute effect subset					0.004
	No intercept	0.347	-0.073	0.767	91
	Yes intercept	0.566	-0.194	1.326	17
	No – Yes contrast	-0.219	-0.979	0.541	
Dominant effect subset					0.000
	No intercept	**0.608**	**0.102**	**1.115**	91
	Yes intercept	0.620	-0.148	1.389	17
	No – Yes contrast	-0.012	-0.699	0.675	

Table S22

Univariate Multilevel Phylogenetic Meta-Regression models with blinding of measurements during judgement bias tests as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint (position of the cue), as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.001
	No intercept	**0.186**	**0.002**	**0.369**	346
	Yes intercept	0.246	-0.009	0.501	113
	No – Yes contrast	-0.061	-0.315	0.194	
Ambiguous cues subset					0.000
	No intercept	0.288	-0.020	0.597	204
	Yes intercept	0.303	-0.091	0.697	65
	No – Yes contrast	-0.015	-0.361	0.331	
Mid-cue effect subset					0.001
	No intercept	0.279	-0.082	0.640	80
	Yes intercept	0.329	-0.139	0.797	28
	No – Yes contrast	-0.05	-0.472	0.373	
Largest absolute effect subset					0.000
	No intercept	0.381	-0.080	0.842	80
	Yes intercept	0.366	-0.253	0.985	28
	No – Yes contrast	0.015	-0.598	0.627	
Dominant effect subset					0.004
	No intercept	**0.652**	**0.143**	**1.161**	80

	Yes intercept	0.488	-0.151	1.127	28
	No – Yes contrast	0.164	-0.403	0.730	

Table S23

Multivariate Multilevel Phylogenetic Meta-Regression models with four moderators that were deemed as significant in the univariate models (Tables S7-S23) as fixed effects. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.072
	Females Active Choice Acoustic Cue intercept	0.028	-0.481	0.536	
	Sex: Female – Male contrast	0.194	-0.102	0.490	
	Sex: Female – Mixed-sex contrast	0.115	-0.160	0.391	
	Task type: Active choice – Go/no-go contrast	0.080	-0.412	0.572	
	Cue type: Acoustic – Olfactory contrast	0.037	-0.585	0.658	
	Cue type: Acoustic – Spatial contrast	-0.103	-0.497	0.292	
	Cue type: Acoustic – Tactile contrast	0.342	-0.302	0.986	
	Cue type: Acoustic – Visual contrast	-0.242	-0.636	0.152	
	Reinforcement type: Reward-Null – Reward-Punishment	0.136	-0.147	0.420	
	Reinforcement type: Reward-Null – Reward-Reward	0.356	-0.161	0.873	
Ambiguous cues subset					0.136
	Females Active Choice Acoustic Cue intercept	0.320	-0.366	1.007	
	Sex: Female – Male contrast	0.295	-0.083	0.673	
	Sex: Female – Mixed-sex contrast	0.196	-0.163	0.555	
	Task type: Active choice – Go/no-go contrast	-0.254	-0.934	0.426	
	Cue type: Acoustic – Olfactory contrast	0.246	-0.600	1.092	
	Cue type: Acoustic – Spatial contrast	-0.163	-0.702	0.376	
	Cue type: Acoustic – Tactile contrast	0.050	-0.721	0.820	
	Cue type: Acoustic – Visual contrast	-0.250	-0.789	0.289	
	Reinforcement type: Reward-Null – Reward-Punishment	0.198	-0.169	0.565	
	Reinforcement type: Reward-Null – Reward-Reward	0.245	-0.432	0.921	
Mid-cue effect subset					0.200
	Females Active Choice Acoustic Cue intercept	0.265	-0.575	1.104	
	Sex: Female – Male contrast	0.360	-0.111	0.832	
	Sex: Female – Mixed-sex contrast	0.252	-0.201	0.705	

	Task type: Active choice – Go/no-go contrast	-0.295	-1.099	0.508	
	Cue type: Acoustic – Olfactory contrast	0.347	-0.714	1.408	
	Cue type: Acoustic – Spatial contrast	-0.229	-0.876	0.417	
	Cue type: Acoustic – Tactile contrast	0.270	-0.649	1.190	
	Cue type: Acoustic – Visual contrast	-0.196	-0.845	0.454	
	Reinforcement type: Reward-Null – Reward-Punishment	0.272	-0.201	0.745	
	Reinforcement type: Reward-Null – Reward-Reward	0.070	-0.774	0.914	
Largest absolute effect subset					0.126
	Females Active Choice Acoustic Cue intercept	-0.327	-1.581	0.927	
	Sex: Female – Male contrast	0.597	-0.118	1.312	
	Sex: Female – Mixed-sex contrast	0.190	-0.516	0.896	
	Task type: Active choice – Go/no-go contrast	0.648	-0.542	1.838	
	Cue type: Acoustic – Olfactory contrast	-0.075	-1.688	1.539	
	Cue type: Acoustic – Spatial contrast	-0.605	-1.596	0.385	
	Cue type: Acoustic – Tactile contrast	0.813	-0.563	2.189	
	Cue type: Acoustic – Visual contrast	-0.452	-1.459	0.555	
	Reinforcement type: Reward-Null – Reward-Punishment	0.409	-0.314	1.132	
	Reinforcement type: Reward-Null – Reward-Reward	0.285	-0.978	1.548	
Dominant effect subset					0.138
	Females Active Choice Acoustic Cue intercept	0.158	-1.004	1.320	
	Sex: Female – Male contrast	0.269	-0.410	0.947	
	Sex: Female – Mixed-sex contrast	0.093	-0.556	0.743	
	Task type: Active choice – Go/no-go contrast	-0.049	-1.142	1.044	
	Cue type: Acoustic – Olfactory contrast	0.887	-0.591	2.366	
	Cue type: Acoustic – Spatial contrast	-0.202	-1.082	0.677	
	Cue type: Acoustic – Tactile contrast	0.873	-0.399	2.145	
	Cue type: Acoustic – Visual contrast	-0.150	-1.056	0.755	
	Reinforcement type: Reward-Null – Reward-Punishment	0.602	-0.049	1.254	
	Reinforcement type: Reward-Null – Reward-Reward	0.008	-1.160	1.176	

Table S24

Results of multivariate meta-regression model selection for the full data set. The top 6 models (out of 16 considered) within the ΔAIC difference (delta) of less than 2 are shown. The models were built from combinations of 4 potentially influential moderators: task type, reinforcement type, cue type and sex of animals. K = the number of parameters in the model including the intercept and the residual error estimates, LogLik = Log Likelihood, AICc = Akaike Information Criteria with correction for small sample sizes, weight = model weights.

Model	K	LogLik	AICc	delta	weight
Task type	6	-581.29	1174.76	0.000	0.346
Reinforcement Category	7	-580.66	1175.56	0.802	0.232
Sex + Task type	8	-580.07	1176.46	1.699	0.148
Reinforcement Category + Task type	8	-580.14	1176.61	1.845	0.138
Intercept-only (no moderators)	5	-583.25	1176.63	1.864	0.136

Table S25

Univariate Multilevel Phylogenetic Meta-Regression models with (scaled) year of study publication as a fixed effect. Intercept values are shown for all levels of categorical moderators and contrasts values are shown for the pairwise differences among these levels. Models include ArticleID, ExperimentID, and ScalePoint, as random effects. CI.lb = lower bound of 95% Confidence Interval; CI.lb = upper bound of 95% Confidence Interval. Bold font indicates effects with Confidence Intervals (CI) not crossing zero (considered statistically significant). R^2 = variance explained ($R^2_{[marginal]}$), k = number of effect sizes per each level of the moderator.

Data	Levels and contrasts	Fixed effects			R^2
		Mean	CI.lb	CI.ub	k
All data					0.000
	Intercept	0.201	0.028	0.375	459
	Publication year slope	-0.002	-0.121	0.118	
Ambiguous cues subset					0.000
	Intercept	0.291	-0.003	0.586	269
	Publication year slope	0.011	-0.143	0.165	
Mid-cue effect subset					0.009
	Intercept	0.295	-0.055	0.644	108
	Publication year slope	0.068	-0.117	0.253	
Largest absolute effect subset					0.009
	Intercept	0.378	-0.024	0.779	108
	Publication year slope	-0.119	-0.407	0.169	
Dominant effect subset					0.000
	Intercept	0.201	0.028	0.375	108
	Publication year slope	-0.002	-0.121	0.118	

Table S26

Bayesian Multivariate Multilevel Phylogenetic Meta-Regression model with four moderators that were deemed as significant in the univariate models with these moderators as a fixed effect. Models include ArticleID, ExperimentID, and ScalePoint (position of the cue), as random effects. Bold font indicates effects with Highest Posterior Density (HPD, i.e. 95% Credible Interval) interval not crossing zero (considered statistically significant). HPD.lb – lower bound of Highest

Posterior Density interval; HPD.lb = upper bound of Highest Posterior Density interval. SD = Standard Deviation, DIC = Deviance Information Criteria.

Fixed effects					
	Mode	Mean	SD	HPD.lb	HPD.ub
Intercept	**0.217**	**0.206**	**0.086**	**0.041**	**0.383**

Random effects					
	Mode	Mean	SD	HPD.lb	HPD.ub
Scale Point	0.003	0.019	0.072	0.000	0.079
ArticleID	0.001	0.035	0.038	0.000	0.107
ExperimentID	0.000	0.030	0.033	0.000	0.098
residuals	0.405	0.416	0.042	0.339	0.501

Heterogeneity				DIC	DIC
	Mode	Mean	SD	Mean	Range
$I^2_{\text{ScalePoint}}$	0.4	3.1	10.9	1037.90	(1037.73 - 1038.15)
$I^2_{\text{ArticleID}}$	0.1	5.4	5.4		
$I^2_{\text{ExperimentID}}$	0.1	4.7	4.8		
I^2_{EffectID}	68.2	66.7	5.3		
I^2_{total}	75.9	76.8	2.0		