Supplementary Materials

SM1: Likelihoods do not necessarily track posterior probabilities in causal-based categorization.

While work in the GM framework assumes that people's judgments track the likelihood $P(o_f|k)$, Rehder has argued that the assumption is consistent with a Bayesian analysis of categorization because a model that computes likelihoods is essentially equivalent to a model that computes the full posterior probabilities (see Rehder, 2015).

Specifically, Rehder (2015) argues that likelihoods are monotonically related to posterior probabilities. That is, if a change in the parameters of the experiment increases the likelihood $P(o_f|k)$, it also increases the posterior probability $P(k|o_f)$, and vice versa. To see this, consider that the posterior probability can be written as[1]:

$$P(k|o_f) = \frac{p(o_f|k)P(k)}{P(o_f|k)P(k) + P(o_f|\neg k)P(\neg k)} \quad (1)$$

where $P(o_f|\neg k)$ is the probability of observing f given that the object does not belong to k, and $p(\neg k) = 1 - p(k)$. These quantities do not depend on the values of the parameters *c*, *m*, *b* that determine the causal model of category k. Therefore, we can replace these quantities with a constant and write $p(o_f|\neg k)p(\neg k) = D$, from which we get:

$$P(k|o_f) = \frac{P(o_f|k)P(k)}{P(o_f|k)P(k) + D} \quad (2)$$

It follows that $P(k|o_f)$ and $P(o_f|k)$ are monotonically related (Rehder, 2015). However, the above proof only shows that $P(k|o_f)$ and $P(o_f|k)$ are monotonically related *with respect to changes to parameters c, b, and m*. That is, the proof shows that if we change (for example) the value of *m*, this will change the likelihood in the same direction that it changes the posterior probability. So, in an experiment

[1] This follows from Bayes' rule (eq.1 main text), and expansion of the denominator according to the law of total probability.

where we only vary *c*, *b*, and *m*, we can expect variations in the likelihood to be a reasonable approximation of variations in the posterior.

Importantly, however, the proof does not apply to changes in other parameters of an experiment. Consider the impact of manipulations of f, the observed features of the object. For instance, in our experiments, participants see objects with features A and B, objects with features ¬A and B, A and ¬B and ¬A and ¬B. The quantity $p(o_f|\neg k)p(\neg k)$ is clearly not invariant to changes in f, so we cannot replace it with a constant if we want to model participants' judgments in an experiment that manipulates the features that participants observe.

As such, observing a given feature (compared to observing its absence) can increase the likelihood while decreasing the posterior probability, or vice-versa. As a simple example, consider a feature f_1 whose probability is .6 in category k but has probability .9 in members of other categories. Given membership in k, observing f_1 is more likely than not observing f_1 (since .6 > .5), therefore $p(f_1|k) > p(\neg f_1|k)$; but observing f_1 might also decrease the probability that the individual belongs to k (since f_1 is more prevalent in other categories), so it is possible that $p(k|f_1) < p(k|\neg f_1)$.

SM2: Robustness across "other categories" causal model parameters

In the main text, to compute posterior probabilities we assume fixed values for the "other categories" causal model parameters (specifically, we set *qc* = .1, *qm* = .2, *qb* = .1, *P(k)* = .2). Here we check whether our results are robust across parameter values, by systematically manipulating each parameter from .1 to .5 in .1 increments (or, for *qm*, from 0 to .99 in .2 increments). For each parameter combination, we computed the fit of each of the three models (we fit the other parameters, i.e., those describing the causal model for category *k*, to the data).

Below we present results obtained using two different measures of model fit. First, the root mean square error (RMSE) is the most demanding measure of model fit for our account: the context-dependent model mostly can achieve a good RMSE fit to the extent that participants use the same response function from probabilities to likert ratings for both consistency and categorization judgments[2].

[2] If this is not the case – for example if participants tend to map a likelihood of .2 to a likert rating of 3 for consistency judgments, but map a posterior probability of .2 to a likert rating of 1 for category membership judgments, then the context-dependent model cannot achieve perfect fit (even in principle) if it assumes that the response function is similar across judgment types. While we do assume (by our use of a single γ parameter) that participants use the same response function

Then, we use a measure of model fit that is not sensitive to whether participants use the same response function across judgment types. Specifically, we compute the correlation between model predictions and average human ratings, separately for consistency and categorization judgments (and then take the average of these two values).

We find that the context-dependent model has the best fit to the data, assessed by RMSE, when qc and *p(k)* are low[3]. When assessing model fit using correlations, we find similar patterns, although the portion of parameter space where the model is best-fitting is even larger.

We also find that the context-dependent model gives the best account of the data for most values of *qm* (the causal strength of the A→ B relationship outside the category), although when *qm* is very high the likelihood model fits best. The analyses below assume the same value of *qm* across different experimental conditions (whereas m is allowed to vary from one condition to the next to account for experimental manipulations of m). We also performed an analysis where we posit that *qm=m* (i.e. that participants assume that the causal strength is always identical outside and inside the category – a corollary is that *qm* covaries with m in response to experimental manipulations), while setting other parameters to the values we use in the main text (*qc* = .1, *qb* = .1, *p(k)* = .2). We find that under, this assumption, the context-dependent model performs less well (RMSE=.826) than the likelihood model (RMSE=.574)[4].

As such, the success of our account relies on the assumption that people think the causal strength of the relationship between A and B is likely to be higher inside compared to outside the category (although it does not require that participants think *qm* is very low). We think this is a reasonable assumption in the context of our experimental materials, which introduce novel causal relationships. Participants probably have no strong priors about whether the causal relationship strength is uniform or variable across categories, and a likely hypothesis in this context is that causal strength is either equally strong across categories, or less strong outside the category. The possibility that the causal relationship might be

---

across judgment types, we do so mostly to make comparison with the other models easier, but we are not committed to the psychological reality of this assumption.

[3] Posterior probabilities increase with p(k), while likelihoods are invariant to p(k). Thus, for high values of p(k) the context-dependent model assigns higher ratings to category membership compared to consistency judgments, which makes it underperform under RMSE evaluation.

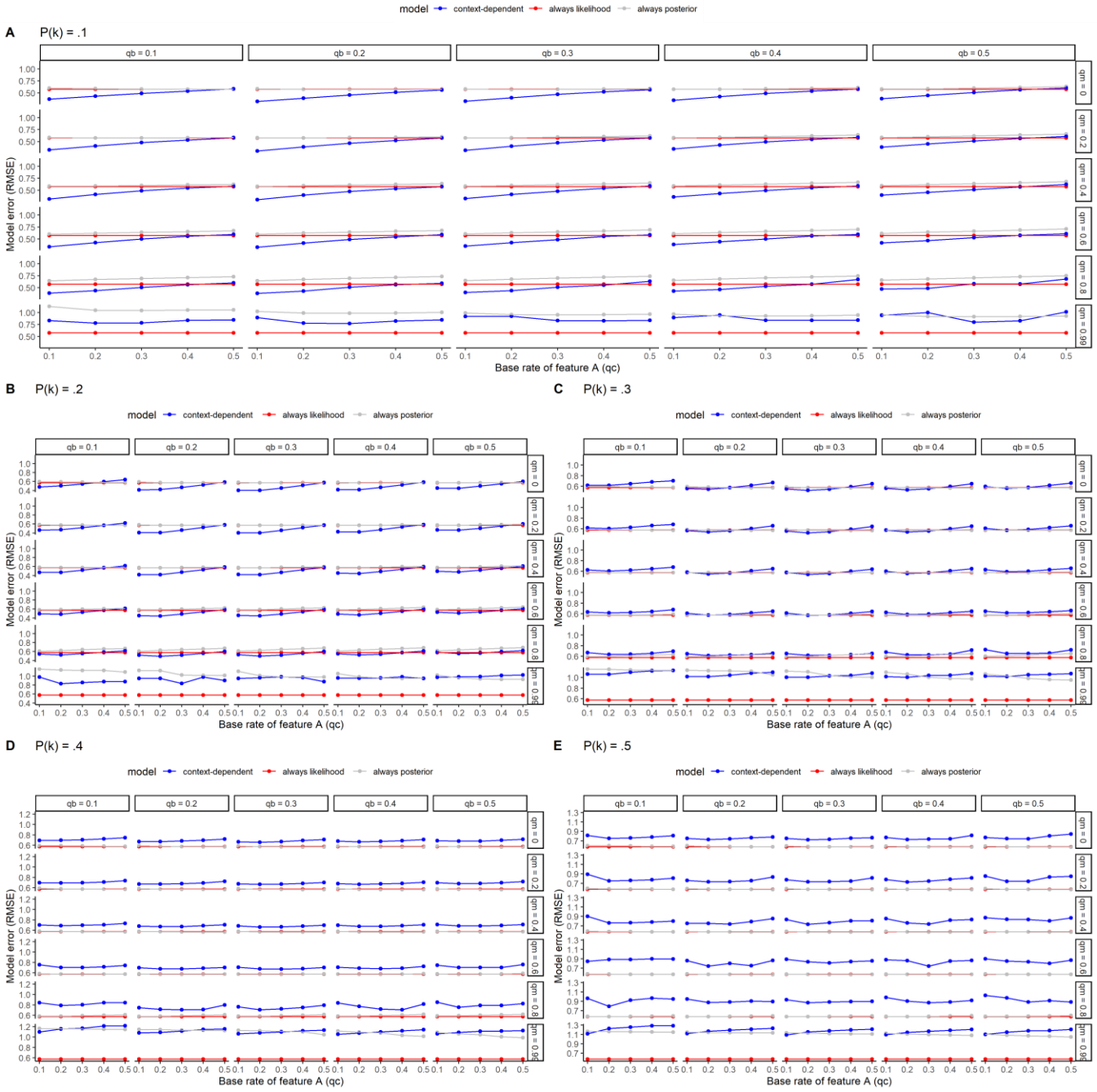
[4] The posterior model performs worst with RMSE=.960.

stronger *outside* the category is relatively implausible, given the sparsity assumption. A Bayesian reasoner averaging across these possibilities should compute an expected value for *qm* that is larger than 0 but lower than m.

In sum, our analyses reveals that the most important assumption for the context-dependent model's adequacy is that *qc* is low, suggesting that feature A is rare outside of category *k*, and that $qm < m$, suggesting that the causal relationship between A and B is likely to be stronger inside than outside category *k*.

Figure S1

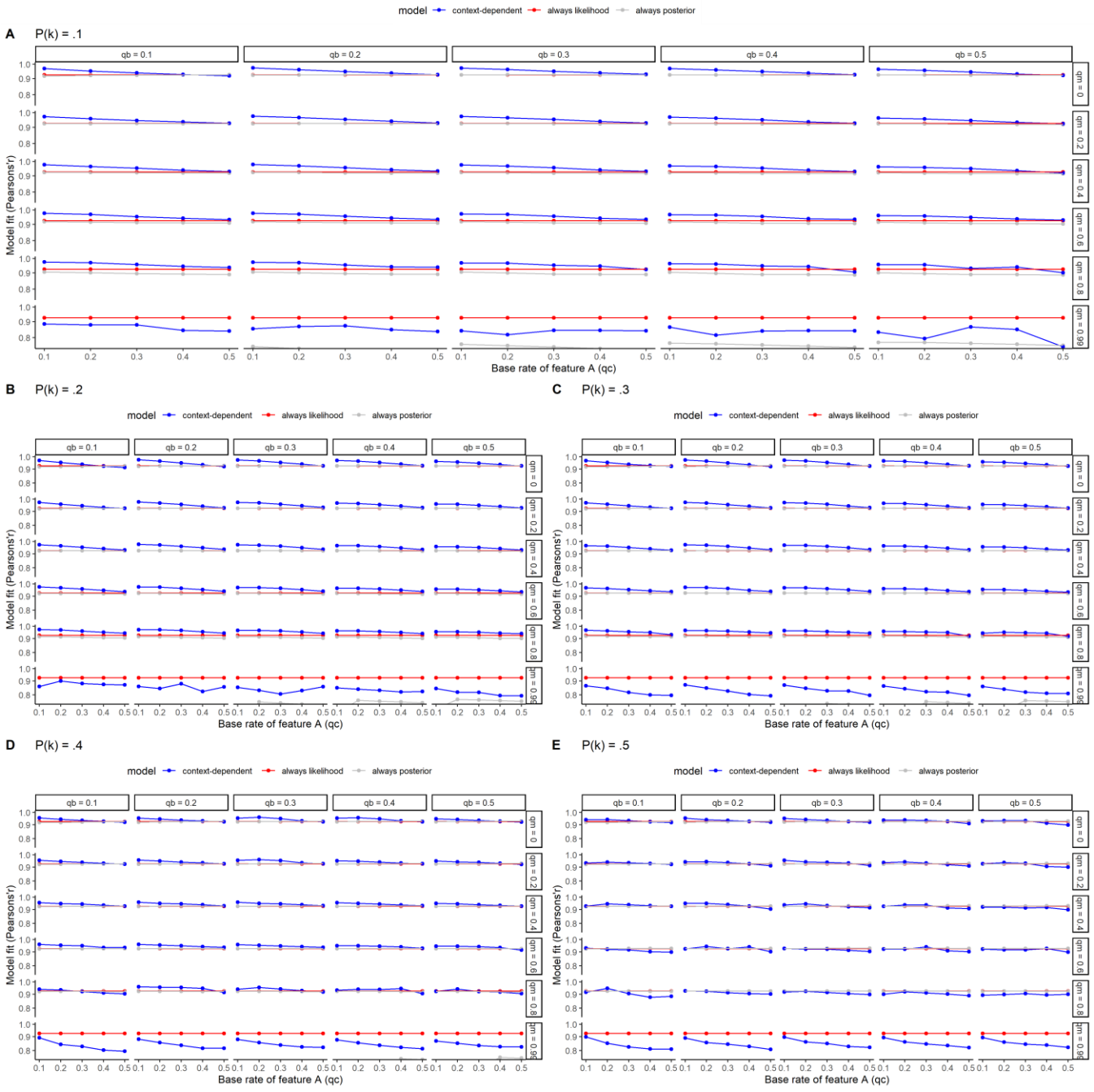
Error for each model, as a function of the "other-categories" causal model parameters.



Note. Through panel (A) to (E) different values of *P(k)* were computed (*P(k)* in {.1, .2, .3, .4, .5). *RMSE*: Root mean squared error. Lower error values represent better model fits.

Figure S2

Model fit for each model, as a function of the "other-categories" causal model parameters.



Note. Model fit is computed as the average of the correlation between model predictions and average participant judgments for each type of judgments (consistency and category membership). Note that unlike in the previous figure, here higher values represent better model fit. Through panel (A) to (E) different values of *P(k)* were computed (*P(k)* in {.1, .2, .3, .4, .5).