



# Is it a judgment of representativeness? Re-examining the birth sequence problem

Lim M. Leong<sup>1</sup> · Johannes Müller-Trede<sup>2</sup> · Craig R. M. McKenzie<sup>1,3</sup>

Accepted: 12 September 2022 / Published online: 11 October 2022  
© The Psychonomic Society, Inc. 2022

## Abstract

Although all birth orders in the “birth sequence problem” are equiprobable, most participants judge the less representative order as less likely than the more representative order. But this well-known problem confounds representativeness with the direction in which birth orders are compared. We hypothesized and corroborated in three experiments (total  $N = 1,136$ ) that participants pragmatically infer the birth orders’ relative prevalence from the direction of comparison. Experiment 1 found that participants judged the less representative sequence as *more* common when we reversed the comparison. Experiment 2 reproduced these results despite removing representativeness as a cue. In Experiment 3, participants preferred to place the relatively common sequence as the referent in an inverted “speaker” problem. Our results turn the iconic problem’s interpretation on its head: Rather than indicating flawed human cognition, the birth sequence problem illustrates people’s ability to adaptively extract subtle linguistic meaning beyond the literal content.

**Keywords** Biases · Cognitive reference points · Heuristics · Judgment · Perceptions of randomness · Representativeness

## Introduction

In making judgments under uncertainty, people are thought to rely on heuristics, or mental rules-of-thumb (Tversky & Kahneman, 1974). Among the most well-known heuristics is representativeness, which holds that people judge subjective probability “by the degree to which [an event or sample] is: (i) similar in essential properties to its parent population; and (ii) reflects the salient features of the process by which it is generated” (Kahneman & Tversky, 1972, p. 431). That is, people base their judgments on the degree to which the target under consideration is a priori similar to a fixed population. The representativeness heuristic has been implicated in biases such as base-rate neglect (Kahneman & Tversky, 1973) and the conjunction fallacy (Tversky &

Kahneman, 1983), and in misperceptions of randomness (for a review, see Nickerson, 2002).

The “birth sequence problem” (Kahneman & Tversky, 1972, p. 432) is an iconic demonstration of representativeness, which asks participants to imagine that:

All families of six children in a city were surveyed. In 72 families the *exact order* of births of boys and girls was GBGBBG. What is your estimate of the number of families surveyed in which the *exact order* of births was BGBBBB?

If births are independent and the probability of a girl (or boy) is 50%, the two birth orders are equally likely. In fact, all exact orders of six children are equiprobable.

But most participants in this study (75/92, or 81.5%) provided an estimate lower than 72, implying that they judged BGBBBB as less common than GBGBBG. The latter sequence with half boys and half girls resembles the general population in terms of the proportions of boys and girls (i.e., approximately 50% of each; James, 1987), while the former sequence with too many boys does not. Kahneman and Tversky (1972) argued that this apparent difference in representativeness is what led participants to judge the latter sequence as more likely. Representativeness can also explain why in a variant of the problem with equal proportions,

✉ Lim M. Leong  
lmleong@ucsd.edu

<sup>1</sup> Department of Psychology, University of California at San Diego, San Diego, 9500 Gilman Drive, #0109, La Jolla, CA 92093, USA

<sup>2</sup> IESE Business School, Barcelona, Spain

<sup>3</sup> Rady School of Management, University of California at San Diego, San Diego, CA, USA

participants judged BBBGGG as less likely than GBBGBG: Repetitions do not reflect the salient features of a random generating process like frequent alternations do. Because the birth sequence problem is so simple and compelling, it has become a mainstay for disseminating judgment and decision research to the wider public (e.g., Hastie & Dawes, 2001; Kahneman, 2011; Lewis, 2016).

In the present article, we propose and corroborate an alternative explanation for this well-known finding. Across three experiments, we show that likelihood judgments in the birth sequence problem cannot be explained by how representative the different birth sequences are of the general population. Instead, the judgments are best explained by conversational pragmatics and cognitive reference points (Gleitman et al., 1996; Rosch, 1975).

### Conversational pragmatics in the birth sequence problem

Research on categorization and cognitive reference points has long recognized that in natural language, variants tend to be placed as subjects, and reference points – items that are more prominent, important, or typical – tend to be placed as complements (Rosch, 1975; see also Wertheimer, 1938). Moreover, these pragmatic tendencies allow people to infer the relative importance or typicality of different items from their syntactic positions. For instance, participants who read sentences with made-up words such as “The zum met the gax” inferred that the item in the complement position (e.g., “gax”) is larger and more important than the one in the subject position (e.g., “zum”; Gleitman et al., 1996). And in a recent study, participants inferred from “Girls do as well at math as boys” that boys (the complement) are more naturally skilled at math than girls (the subject), even though the sentence explicitly expresses equality (Chestnut & Markman, 2018).

Similar inferences may be drawn from the comparative structure in the birth sequence problem, where participants are first informed about the prevalence of a birth order and then asked to estimate the prevalence of another. Extending Rosch’s (1975) theory to this problem, we propose that the initial sequence serves as the referent (complement), and the sequence to be estimated serves as the target (subject). In line with prior studies on pragmatic inferences, participants may then infer that the target sequence is less typical or common than the referent. That is, the directionality of the comparison between sequences, rather than the sequences’ representativeness, could determine participants’ likelihood judgments.

Importantly, Kahneman and Tversky (1972) neglected to test whether their finding holds if the direction of comparison is reversed. In other words, their original experiment confounds representativeness with the directionality of the

comparison. If participants are first told that 72 families have the *less* representative birth order BBBBBB, will they then provide estimates *higher* than 72 for the *more* representative birth order GBGBBG?

To this end, Experiment 1 provides a conceptual replication with a simple twist: We inverted the direction of comparison and placed the less representative sequence as the referent.

Because this manipulation leaves the sequences and their key characteristics (i.e., the proportions of boys and girls, and the alternations between the genders) untouched, how representative they are of the general population remains the same. The representativeness heuristic therefore predicts that regardless of the direction of comparison, participants should judge the less representative sequence as less common. In contrast, conversational pragmatics predicts that participants should judge the target sequence, regardless of its representativeness, as less common. Crucially, this implies the opposite prediction: When we reverse the direction of comparison, participants should judge the less representative sequence as *more* common.

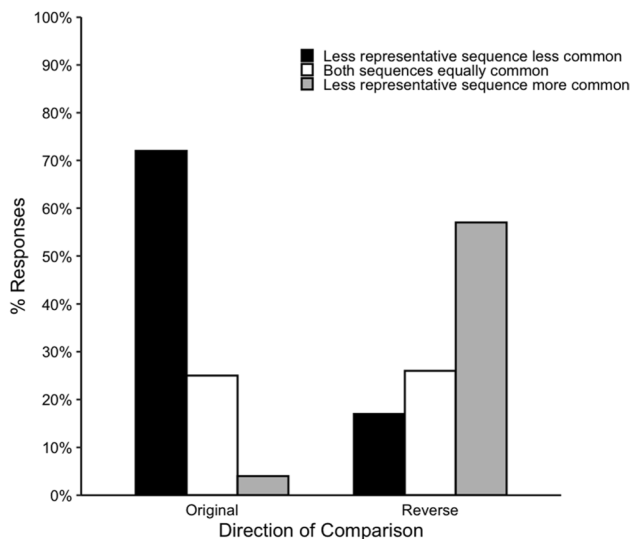
### Experiment 1

Experiment 1 tests whether the direction of comparison affects likelihood judgments in the birth sequence problem. It includes direct replications of Kahneman and Tversky (1972), in which the more representative sequence served as the referent, and conceptual replications, in which the less representative sequence served as the referent.

### Method

Participants were 388 University of California at San Diego (UCSD) undergraduate students ( $M_{\text{age}} = 20.0$  years, one participant did not report age; 71% female) who received partial course credit. In all experiments reported in this article, we used a convenience sample and recruited participants for the duration of an academic term. Participants who completed any one of our experiments were barred from signing up for all subsequent ones.

We employed Kahneman and Tversky’s (1972) birth sequence problem using their exact wording found in our *Introduction*, and tested both the version with sequences of unequal proportions and the version with sequences of equal proportions. For each version, we manipulated the direction of comparison. In the “original comparison” conditions, participants read that 72 families have the more representative birth sequence and were asked to estimate the less representative birth sequence. In the “reverse comparison” conditions, participants instead read that 72 families have the *less* representative birth sequence and were



**Fig. 1** Experiment 1 results. Percentage of responses implying that the less representative sequence was less common than, equally as common as, or more common than the more representative sequence, in the original and reverse direction of comparison

asked to estimate the *more* representative birth sequence. Participants were randomly assigned to one of four between-subjects conditions and provided their estimates by typing in a numerical value.

## Results

We categorized the responses according to whether they implied the less representative sequence (BGBBBB in the unequal- or BBBGGG in the equal-proportions version of the problem) was less common than, equally as common as, or more common than the more representative sequence (GBGBBG in the unequal- or GBBGBG in the equal-proportions version of the problem). Results were similar for the unequal- and equal-proportions versions of the problem, and a log-linear analysis found the problem version not to be significantly associated with responses or direction of comparison,  $G^2(5, N = 388) = 9.59, p = .088$ . The following analyses thus collapse across the two versions of the problem.<sup>1</sup>

Figure 1 reveals that, as predicted by conversational pragmatics but not by the representativeness heuristic, there was a significant direction-of-comparison effect,  $\chi^2(2, N = 388) = 156.22, p < .001$ . In the original comparison conditions, where the more representative sequence served as the referent, 71.5% (138/193) of participants judged the less representative sequence as less common (sign test,  $p <$

.001). We thus replicated the original finding by Kahneman and Tversky (1972). But in the reverse comparison conditions, where the less representative sequence served as the referent, only 16.9% (33/195) of participants judged the less representative sequence as less common (sign test,  $p < .001$ ). Instead, 56.9% (111/195) of participants judged the less representative sequence as *more* common (sign test,  $p < .001$ ), for an odds ratio of 6.49, 95% confidence interval (CI) [4.06, 10.37], in favor of conversational pragmatics over the representativeness heuristic.<sup>2</sup> In the Online Supplementary Materials (OSM), we further show that this direction-of-comparison effect generalizes from birth orders in families of six to coin flips in series of four (Experiment S1) and from the reference value of 72 to a lower reference value of 12 (Experiment S2).

## Experiment 2

Experiment 2 introduces a novel variant of the birth sequence problem that retains its comparison structure but uses stimuli that do not differ in their representativeness. Removing representativeness as a cue allows us to investigate its contribution to the original finding that people judge the target as less likely than the referent. In the novel variant, a representativeness heuristic no longer predicts a bias, whereas conversational pragmatics predict that participants would continue to judge the target as less likely than the referent.

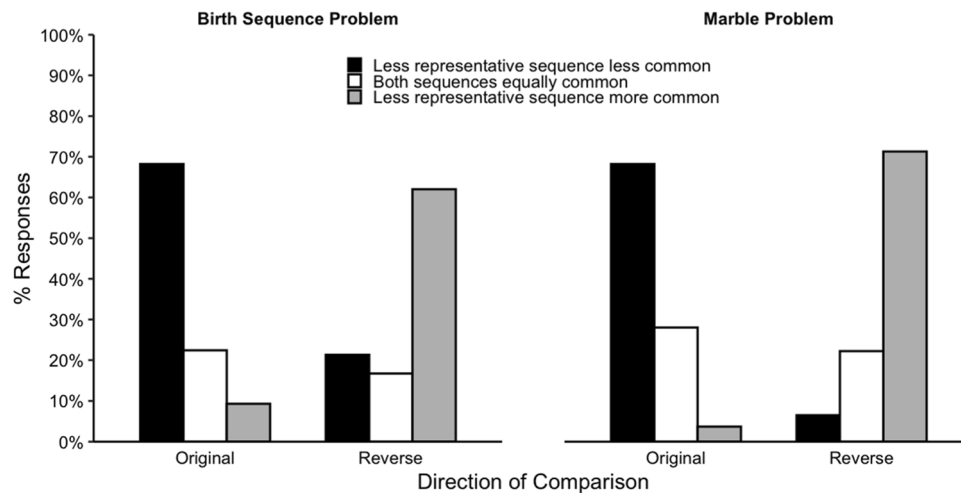
## Method

We included an attention check and a self-report measure for whether participants had previously seen the problem (see below). After excluding 84 participants who either failed the attention check or reported having previously seen the problem, we were left with a final sample of 430 UCSD undergraduate students ( $M_{\text{age}} = 21.28$  years; 63.3% female, one participant reported “other” and three reported “prefer not to say”) who participated for partial course credit.

We manipulated the direction of comparison in two different versions of the problem. In the birth sequence problem, participants read the standard problem with

<sup>1</sup> See Online Supplementary Materials (OSM) for detailed breakdowns for each version separately.

<sup>2</sup> We focus on the odds ratio in the reverse comparison conditions because only in those conditions do pragmatics and representativeness make divergent predictions. The two accounts’ overall empirical performance may be measured by including both the original and the reverse comparison conditions. In that analysis, conversational pragmatics can account for 64.2% of responses (249/388), whereas the representativeness heuristic can account for only 44.1% (171/388), with an odds ratio of 2.27 in favor of the former, 95% CI [1.70, 3.03].



**Fig. 2** Experiment 2 results. Percentage of responses implying that BBBBBB (left panel) was, or blue marbles (right panel) were, less common than, equally as common as, or more common than GBGBBG or red marbles, as a function of problem and direction of com-

parison. In the marble problem, red marbles served as the referent in the “original comparison” condition, and blue marbles served as the referent in the “reverse comparison” condition

sequences that differed in terms of representativeness. In the “marble problem,” participants instead read the following:

Imagine an assortment of marbles of various colors. Marbles of each color were counted. 72 of the marbles were red (blue). What is your estimate of the number of marbles that were blue (red)?

Because the distribution and frequency of marble colors are not given, there is no way to determine the “correct” answer to this problem. Furthermore, red marbles are presumably not any more or less representative than blue marbles of an assortment of marbles of unknown colors, so participants could not base their estimates on representativeness.

Participants were randomly assigned to one of four between-subjects conditions (problem  $\times$  direction of comparison) and provided their estimates by typing in a numerical value. Afterward, to check whether they were paying attention, participants were asked “The problem you just answered mentioned which of the following?” and could select among the choices “Families with different birth orders,” “Marbles of various colors,” and “None of the above.”

## Results

We again categorized the responses in the birth sequence problem according to whether they implied the less representative sequence was less common than, equally as common as, or more common than the more representative sequence. For the marble problem, we arbitrarily designated the condition in which red marbles serve as the referent as the “original

comparison,” and the condition in which blue marbles serve as the referent as the “reverse comparison.” Our results and their statistical significance remain qualitatively unchanged if we instead designate the condition in which blue marbles serve as the referent as the “original comparison.” A log-linear analysis found that problem (birth sequence vs. marble), direction of comparison (original vs. reverse), and responses (less common vs. equally common vs. more common) were significantly associated with each other,  $G^2(2, N = 430) = 10.18, p = .006$ . We therefore discuss the results separately for the birth sequence and marble problem.

Replicating our results from Experiment 1, we found a large direction-of-comparison effect in the birth sequence problem,  $\chi^2(2, N = 215) = 69.09, p < .001$  (Fig. 2, left panel). In the original comparison condition, 68.2% (73/107) of participants judged the less representative sequence as less common (sign test,  $p < .001$ ), but in the reverse comparison condition, only 21.3% (23/108) of participants did (sign test,  $p < .001$ ). As in Experiment 1, most participants (62.0% or 67/108) in the reverse comparison condition judged the less representative sequence as *more common* (sign test,  $p < .001$ ), for an odds ratio of 6.04, 95% CI [3.31, 11.03], in favor of conversational pragmatics over the representativeness heuristic.

Strikingly, Fig. 2 reveals a very similar pattern of results for the marble problem (right panel). Although representativeness is not well-defined in this problem, we again find a large direction-of-comparison effect,  $\chi^2(2, N = 215) = 120.90, p < .001$ . In the original comparison condition, where red marbles were the referent, 68.2% (73/107) of participants judged blue marbles as less common (sign test,  $p < .001$ ). But in the reverse comparison condition, where blue marbles were the referent, only 6.5% (7/108) of participants

did (sign test,  $p < .001$ ). As predicted by the pragmatics account, most participants (71.3% or 77/108) instead judged blue marbles as *more* common when it was the referent (sign test,  $p < .001$ ).

We also examined the differences across the birth sequence and marble problem for each direction of comparison separately. For the original direction of comparison, we did not observe a significant difference between the two problems,  $\chi^2(2, N = 214) = 3.24, p = .20$ . That is, when the more representative sequence acted as the referent, participants responded to the marble problem much like how they responded to the birth sequence problem. In contrast, for the reverse direction of comparison, responses differed significantly across the two problems,  $\chi^2(2, N = 216) = 10.09, p = .006$ . When the less representative sequence acted as the referent in the birth sequence problem, participants responded to it differently compared to how they responded to the referent in the marble problem, and more frequently judged it to be less common (21.3%, or 23/108, vs. 6.5%, or 7/108). This small asymmetry suggests that although representativeness fails to explain the key features of the data, it does play a minor role in participants' likelihood judgments. We return to this observation in the *General discussion*.

Overall, Experiment 2 overwhelmingly favors the conversational pragmatics account. As in Experiment 1, whichever sequence served as the referent was judged to be more common. Furthermore, the marble problem – where representativeness does not make a prediction – yielded virtually identical results. Whether sequences are judged to be relatively common or uncommon thus appears to be largely determined by conversational pragmatics, and not by the sequences' representativeness.

### Experiment 3

In this preregistered experiment, we examined whether the directionality of the comparison indeed signals relative prevalence by reversing the task: We manipulated the two birth orders' prevalence and asked participants which they preferred to place as the referent. This allowed us to assess the adaptiveness of the likelihood judgments in the original birth sequence problem. If participants prefer to place the relatively common sequence as the referent, then the biased likelihood judgments in the original problem reflect an adaptive response to the social environment they are embedded in.

### Method

Participants were 318 UCSD undergraduate students ( $M_{\text{age}} = 19.97$  years; 73.6% female, three participants reported

“other” and two “prefer not to say”) who received partial course credit. They were randomly assigned to one of two conditions. In the “GBGBBG common” condition, participants read the following:

Imagine that you are interested in demography, that is, the study and statistics of human populations. One aspect of demography that you find particularly interesting is the birth order of girls (G) and boys (B) in families. You surveyed the exact birth orders of all families of six children in a particular city, and found that the exact birth order GBGBBG is relatively common, and the exact birth order BGBBBB is relatively uncommon.

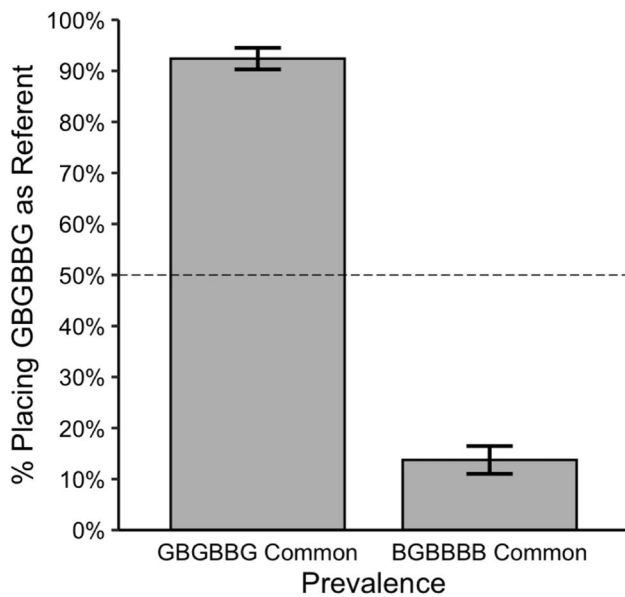
You now want to write up the information you have gathered for a friend who is also interested in birth orders. You start with the following opening: All families of six children in a city were surveyed. In 72 families the exact order of births of boys and girls was \_\_\_\_\_.

The “BGBBBB common” condition was identical, except participants read that BGBBBB is relatively common and GBGBBG is relatively uncommon.

Afterward, they were asked “Given what you know about their relative prevalence, which of the two sequences mentioned above would you refer to in the blank?” Participants chose between “GBGBBG (relatively common)” and “BGBBBB (relatively uncommon)” in the GBGBBG common condition, and between “GBGBBG (relatively uncommon)” and “BGBBBB (relatively common)” in the BGBBBB common condition. The order of presentation for the two options (left vs. right) was counter-balanced.

### Results

Figure 3 shows the percentage of participants who placed GBGBBG in the referent position as a function of prevalence. As predicted, we found that whether participants place GBGBBG as the referent depends on whether it was said to be relatively common,  $\chi^2(1, N = 318) = 194.22, p < .001$ . When GBGBBG was said to be relatively common, 92.4% (146/158) of participants chose to place GBGBBG as the referent (binomial test,  $p < .001$ ). But when BGBBBB was said to be relatively common, only 13.8% (22/160) of participants chose to do so (binomial test,  $p < .001$ ). Participants thus showed a strong preference for placing the relatively common sequence in the referent position regardless of the configuration of boys and girls in each sequence. Given this preference, people are therefore warranted in inferring that the referent sequence in the original problem is more common.



**Fig. 3** Experiment 3 results. Percentage of participants placing GBGBBG in the referent position as a function of prevalence. Error bars represent the standard error of the proportion

## General discussion

Likelihood judgments in the birth sequence problem have traditionally been explained in terms of the representativeness heuristic, with the implication that human perception of chance is fundamentally flawed (Kahneman & Tversky, 1972). In this article, we show that, contrary to appearances, people’s judgments are not driven by representativeness but by conversational pragmatics that adaptively reflect the social environment that they are embedded in (Gleitman et al., 1996; Rosch, 1975).

Experiment 1 showed that likelihood judgments strongly depend on the direction of comparison. When the more representative sequence served as the referent, as in the original experiment, we replicated the finding that participants judge the less representative sequence as less likely. But when the comparison was reversed and the less representative sequence served as the referent, participants judged the less representative sequence as *more* likely. Experiment 2 found a nearly identical pattern of results in a novel marble problem that preserved the comparison structure but eliminated representativeness as a cue. Finally, Experiment 3 placed participants in the role of “speakers” and discovered that they strongly preferred placing the relatively common sequence as the referent, regardless of its representativeness.

Limited evidence for representativeness only emerged in the form of minor asymmetries in likelihood judgments when direction of comparison was manipulated. Across our replications of the birth sequence problem, the dominant tendency to attribute relative prevalence to the referent was

slightly attenuated when the referent was the less representative sequence compared to when it was not (e.g., Fig. 1). Tellingly, the asymmetries disappeared when we removed representativeness as a cue in Experiment 2’s marble problem (Fig. 2, right panel). We suspect that these robust but modest asymmetries may reflect the role of representativeness as a valid cue for inferring whether a sequence was generated by a random process (Griffiths et al., 2018; see also Hahn & Warren, 2009; Miller & Sanjurjo, 2018).

Furthermore, a sizable minority (roughly 25%) of participants across our experiments responded that the two sequences are equally likely, which is the “correct” answer under the standard interpretation of the problem. Their responses do not reflect either conversational pragmatics or representativeness, and perhaps these participants treated the problem like the exercise in probability theory that it was intended as. Curiously, however, a similar proportion of participants judged the two marble colors to be equally likely in the marble problem, although this problem lacks a correct solution. This raises the possibility that even those who gave the “correct” answer in the birth sequence problem may not have reasoned about the probability of random sequences at all.

Could representativeness be salvaged by assuming that participants are inferring the population (or generating process) from the referent in the birth order problem? For example, when the referent in our reverse comparison condition was BGBBBB, might participants have assumed that the sequence was representative of the population it had been sampled from, and that is why they judged it as more likely than the target? We find such an attempt to reconcile our results with the representativeness heuristic to be both implausible and problematic. It is implausible because it assumes that when presented with the referent BGBBBB, participants infer a highly unusual population in which mostly boys are born.<sup>3</sup> It is also problematic because it directly contradicts the logic of Kahneman and Tversky’s (1972) original study, in which they considered it “obvious” that GBGBBG was more representative than BGBBBB. The population was thus assumed to be fixed (i.e., males and females are equally likely and independent), and the sequences’ representativeness was determined a priori based on their similarity to the fixed population. BGBBBB was predicted to be judged less likely in the original study not because it was the target rather than referent, but because its features were less similar to – and thus less representative of – the fixed population. In the

<sup>3</sup> The human sex ratio at birth is essentially constant at approximately equal proportions boys and girls, both at the population level (Chahnazarian, 1988; James, 1987) and for individual mothers and fathers (Zietsch et al., 2020).

logic of the original study, the distinction between referent and target is irrelevant to the representativeness heuristic.

The conversational pragmatics revealed by our experiments, in contrast, are neither implausible nor problematic, but seem consistent with rational models of decision making (e.g., Chater & Oaksford, 1999; Gershman et al., 2015; Griffiths et al., 2015). The pragmatic inferences do not imply that participants presented with an unusual referent like BGBBBB must infer that it has been sampled from an equally unusual population. Instead, the inferences reflect how speakers communicate information about relative prevalence (Experiment 3), and the resulting likelihood judgments appear to be adaptive responses to the social environment that they are embedded in. This finding adds to a growing literature that illustrates people's remarkable ability to extract subtle meaning beyond the literal content of utterances (Krijnen et al., 2017; McKenzie, 2004; McKenzie et al., 2018; McKenzie & Nelson, 2003; Schwarz, 1994; Sher & McKenzie, 2006; Tannenbaum et al., 2013; Wänke & Reuter, 2010). And it turns the conventional interpretation of the birth sequence problem on its head: Rather than indicating flawed human cognition, the problem illustrates people's ability to adaptively extract subtle linguistic meaning beyond the literal content.

The representativeness heuristic has been criticized before. Gigerenzer (1991, 1996), for instance, questioned its descriptive validity and predictive usefulness in the context of other purported biases, such as base-rate neglect and the conjunction fallacy (see also Hertwig & Gigerenzer, 1999; Koehler, 1996). But the birth sequence problem is still widely considered a compelling example of the heuristic and its bleak implications for rationality. For the past 50 years, this problem has been an integral part of a popular narrative on judgment and decision making that combines simple experiments with intriguing claims about human irrationality (Kahneman, 2011; Lewis, 2016). Popular summaries of this research, however, tend to omit the lively debate within psychology about what constitutes the right benchmark for judging behaviors as rational or irrational (e.g., Gershman et al., 2015; Gigerenzer, 1991; Hertwig & Herzog, 2009; Koehler, 1996; McKenzie et al., 2018; Stanovich, 1999). Our findings provide a much-needed, albeit late-in-coming, correction to the traditional interpretation of the problem and illustrate how ostensible biases can sometimes reflect the sophistication of human cognition rather than its shortcomings.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.3758/s13423-022-02188-9>.

**Funding** This research was supported by a National Science Foundation grant (SES-2049935) to C.R.M.M.

**Data availability** Data and materials for all experiments can be found at the Open Science Framework page for this article: [https://osf.io/2m8nk/?view\\_only=d5817c2c5f7a4899a5bd520c4a3bf388](https://osf.io/2m8nk/?view_only=d5817c2c5f7a4899a5bd520c4a3bf388). Experiment 3 was preregistered. We have no conflicts of interest to disclose.

## References

- Chahnazarian, A. (1988). Determinants of the sex ratio at birth: Review of recent literature. *Social Biology*, 35(3-4), 214–235.
- Chater, N., & Oaksford, M. (1999). Ten years of the rational analysis of cognition. *Trends in Cognitive Sciences*, 3(2), 57–65.
- Chestnut, E. K., & Markman, E. M. (2018). “Girls are as good as boys at math” implies that boys are probably better: A study of expressions of gender equality. *Cognitive Science*, 42, 2229–2249.
- Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, 349, 273–278.
- Gigerenzer, G. (1991). How to make cognitive illusions disappear: Beyond “heuristics and biases”. *European Review of Social Psychology*, 2, 83–115.
- Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). *Psychological Review*, 103, 592–596.
- Gleitman, L., Gleitman, H., Miller, C., & Ostrin, R. (1996). Similar, and similar concepts. *Cognition*, 58(3), 321–376.
- Griffiths, T. L., Lieder, F., & Goodman, N. D. (2015). Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2), 217–229.
- Griffiths, T. L., Daniels, D., Austerweil, J. L., & Tenenbaum, J. B. (2018). Subjective randomness as statistical inference. *Cognitive Psychology*, 103, 85–109.
- Hahn, U., & Warren, P. A. (2009). Perceptions of randomness: Why three heads are better than four. *Psychological Review*, 116, 454–461.
- Hastie, R., & Dawes, R. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making*. Sage.
- Hertwig, R., & Gigerenzer, G. (1999). The ‘conjunction fallacy’ revisited: How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making*, 12(4), 275–305.
- Hertwig, R., & Herzog, S. M. (2009). Fast and frugal heuristics: Tools of social rationality. *Social Cognition*, 27(5), 661–698.
- James, W. H. (1987). The human sex ratio. Part I: Review of the literature. *Human Biology*, 59(5), 721–752.
- Kahneman, D. (2011). *Thinking fast and slow*. Farrar, Straus & Giroux.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3, 430–454.
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251.
- Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences*, 19(1), 1–17.
- Krijnen, J. M. T., Tannenbaum, D., & Fox, C. R. (2017). Choice architecture 2.0: Behavioral policy as an implicit social interaction. *Behavioral Science & Policy*, 3(2), 1–18.
- Lewis, M. (2016). *The undoing project: A friendship that changed our minds*. Norton.
- McKenzie, C. R. M. (2004). Framing effects in inference tasks -- and why they are normatively defensible. *Memory and Cognition*, 32, 874–885.
- McKenzie, C. R. M., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin and Review*, 10, 596–602.

- McKenzie, C. R. M., Sher, S., Leong, L. M., & Müller-Trede, J. (2018). Constructed preferences, rationality, and choice architecture. *Review of Behavioral Economics*, *5*, 337–360.
- Miller, J. B., & Sanjurjo, A. (2018). Surprised by the hot hand fallacy? A truth in the law of small numbers. *Econometrica*, *86*(6), 2019–2047.
- Nickerson, R. S. (2002). The production and perception of randomness. *Psychological Review*, *109*, 330–357.
- Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*, 532–547.
- Schwarz, N. (1994). Judgment in a social context: Biases, shortcomings, and the logic of conversation. *Advances in Experimental Social Psychology*, *26*, 123–162.
- Sher, S., & McKenzie, C. R. M. (2006). Information leakage from logically equivalent frames. *Cognition*, *101*, 467–494.
- Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.
- Tannenbaum, D., Valasek, C. J., Knowles, E. D., & Ditto, P. H. (2013). Incentivizing wellness in the workplace: Sticks (not carrots) send stigmatizing signals. *Psychological Science*, *24*(8), 1512–1522.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*, 1124–1131.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293–315.
- Wänke, M., & Reuter, L. (2010). Direction-of-comparison effects: How and why comparing apples with oranges is different from comparing oranges with apples. In G. Keren (Ed.), *Perspectives on Framing* (pp. 177–194). Psychology Press.
- Wertheimer, M. (1938). Numbers and numerical concepts in primitive peoples. In W. D. Ellis (Ed.), *A source book of Gestalt psychology* (pp. 265–273). The Gestalt Journal Press.
- Zietsch, B. P., Walum, H., Lichtenstein, P., Verweij, K. J., & Kuja-Halkola, R. (2020). No genetic contribution to variation in human offspring sex ratio: a total population study of 4.7 million births. *Proceedings of the Royal Society B*, *287*(1921), 20192849.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.