FAIR COINS TEND TO LAND ON THE SAME SIDE THEY STARTED: EVIDENCE FROM 350,757 FLIPS

František Bartoš[1*], Alexandra Sarafoglou[1], Henrik R. Godmann[1], Amir Sahrani[1], David Klein Leunk[2], Pierre Y. Gui[2], David Voss[2], Kaleem Ullah[2], Malte J. Zoubek[3], Franziska Nippold, Frederik Aust[1], Felipe F. Vieira[4], Chris-Gabriel Islam[5,6], Anton J. Zoubek[7], Sara Shabani[8], Jonas Petter[1], Ingeborg B. Roos[9], Adam Finnemann[1,10], Aaron B. Lob[2,11], Madlen F. Hoffstadt[1], Jason Nak[2], Jill de Ron[2], Koen Derks[12], Karoline Huth[2,13], Sjoerd Terpstra[14], Thomas Bastelica[15,16], Magda Matetovici[2,17], Vincent L. Ott[2], Andreea S. Zetea[2], Katharina Karnbach[2], Michelle C. Donzallaz[1], Arne John[2], Roy M. Moore[2], Franziska Assion[1,8], Riet van Bork[2], Theresa E. Leidinger[2], Xiaochang Zhao[2], Adrian Karami Motaghi[2], Ting Pan[1,9], Hannah Armstrong[2], Tianqi Peng[2], Mara Bialas[20], Joyce Y.-C. Pang[2], Bohan Fu[2], Shujun Yang[2], Xiaoyi Lin[2], Dana Sleiffer[2], Miklos Bognar[21], Balazs Aczel[22], and Eric-Jan Wagenmakers[1]

[1]Department of Psychological Methods, University of Amsterdam,
[2]Department of Psychology, University of Amsterdam,
[3]Institute of Economic Law, University of Kassel,
[4]Faculty of Psychology and Educational Sciences, KU Leuven,
[5]Chairs of Statistics and Econometrics, Georg August University of Göttingen,
[6]Centre for Environmental Sciences, Hasselt University,
[7]Gutenberg School Wiesbaden,
[8]Department of Psychology, Justus Liebig University,
[9]Department of Adult Psychiatry, Amsterdam University Medical Centre,
[10]Centre for Urban Mental Health, University of Amsterdam,
[11]Cognitive and Behavioral Decision Research, University of Zurich,
[12]Center of Accounting, Auditing & Control, Nyenrode Business University,
[13]Department of Psychiatry, Amsterdam UMC, University of Amsterdam,
[14]Institute for Marine and Atmospheric Research Utrecht, Utrecht University,
[15]Centre Hospitalier Le Vinatier,
[16]INSERM, U1028; CNRS, UMR5292, Lyon Neuroscience Research Center, Psychiatric Disorders: from Resistance to Response Team,
[17]Research Institute of Child Development and Education, University of Amsterdam,
[18]Faculty of Behavioral and Movement Sciences, Vrije Universiteit Amsterdam,
[19]Department of Develomental Psychology, University of Amsterdam,
[20]Behavioural Science Institute, Radboud University,
[21]Doctoral School of Psychology, ELTE Eotvos Lorand University,
[22]Institute of Psychology, ELTE Eotvos Lorand University

*Correspondence concerning this article should be addressed to František Bartoš at f.bartos96@gmail.com.
For all but the last four positions, the authorship order aligns with the number of coin flips contributed.

ABSTRACT

Many people have flipped coins but few have stopped to ponder the statistical and physical intricacies of the process. In a preregistered study we collected 350,757 coin flips to test the counterintuitive prediction from a physics model of human coin tossing developed by Diaconis, Holmes, and Montgomery (DHM; 2007). The model asserts that when people flip an ordinary coin, it tends to land on the same side it started—DHM estimated the probability of a same-side outcome to be about 51%. Our data lend strong support to this precise prediction: the coins landed on the same side more often than not, $\text{Pr(same side)} = 0.508$, 95% credible interval (CI) [0.506, 0.509], $\text{BF}_{\text{same-side bias}} = 2359$. Furthermore, the data revealed considerable between-people variation in the degree of this same-side bias. Our data also confirmed the generic prediction that when people flip an ordinary coin—with the initial side-up randomly determined—it is equally likely to land heads or tails: $\text{Pr(heads)} = 0.500$, 95% CI [0.498, 0.502], $\text{BF}_{\text{heads-tails bias}} = 0.182$. Furthermore, this lack of heads-tails bias does not appear to vary across coins. Additional exploratory analyses revealed that the within-people same-side bias decreased as more coins were flipped, an effect that is consistent with the possibility that practice makes people flip coins in a less wobbly fashion. Our data therefore provide strong evidence that when some (but not all) people flip a fair coin, it tends to land on the same side it started. Our data provide compelling statistical support for the DHM physics model of coin tossing.

Introduction

A coin flip—the act of spinning a coin into the air with your thumb and then catching it in your hand—is often considered the epitome of a chance event. It features as a ubiquitous example in textbooks on probability theory and statistics [1, 2, 3, 4, 5] and constituted a game of chance ('capita aut navia'–'heads or ships') already in Roman times (6, $\sim$ 431 AD, 1.7:22).

The simplicity and perceived fairness of a coin flip, coupled with the widespread availability of coins, may explain why it is often used to make even high-stakes decisions. For example, a coin flip was used to determine which of the Wright brothers would attempt the first flight in 1903; who would get the last plane seat for the tour of rock star Buddy Holly (which crashed and left no survivors) in 1959; the winner of the European Championship semi-final soccer match between Italy and the Soviet Union (an event which Italy went on to win) in 1968; which of two companies would be awarded a public project in Toronto in 2003; and to break the tie in local political elections in the Philippines in both 2004 and 2013.

Despite the widespread popularity of coin flipping, few people pause to reflect on the notion that the outcome of a coin flip is anything but random: a coin flip obeys the laws of Newtonian physics in a relatively transparent manner [3]. According to the standard model of coin flipping [7, 8, 9, 10, 11], the flip is a deterministic process and the perceived randomness originates from small fluctuations in the initial conditions (regarding starting position, configuration, upward force, and angular momentum) combined with narrow boundaries on the outcome space. Therefore the standard model predicts that when people flip a fair coin, the probability of it landing heads is 50% (i.e., there is no 'heads-tails bias'; conversely, if a coin would land on one side more often than the other, we would say there is a 'heads-tails bias').[1]

The standard model of coin flipping was extended by Diaconis, Holmes, and Montgomery (DHM; [13]) who proposed that when people flip a ordinary coin, they introduce a small degree of 'precession' or wobble—a change in the direction of the axis of rotation throughout the coin's trajectory. According to the DHM model, precession causes the coin to spend more time in the air with the initial side facing up. Consequently, the coin has a higher chance of landing on the same side as it started (i.e., 'same-side bias'). Under the DHM model, this same-side bias is absent only when there is no wobble whatsoever, as any non-zero angle of rotation results in a same-side bias (with a higher degree of wobble resulting in a more pronounced bias). Based on a modest number of empirical observations (featuring coins with ribbons attached and high-frame-rate video recordings) [13] measured the off-axis rotations in typical human flips. Based on these observations, the DHM model predicted that a coin flip should land on the same side as it started with a probability of approximately 51%, just a fraction higher than chance.

Throughout history, several researchers have collected thousands of coin flips. In the 18$^{\text{th}}$ century, the famed naturalist Count de Buffon [14] collected 2,048 uninterrupted sequences of 'heads' in what is possibly the first statistical experiment ever conducted. In the 19$^{\text{th}}$ century, the statistician Karl Pearson [15] flipped a coin 24,000 times to obtain 12,012 tails. And in the 20$^{\text{th}}$ century, the mathematician John Kerrich [1] flipped a coin 10,000 times for a total of 5,067 heads while interned in Nazi-occupied Denmark. These experiments do not allow a test of the DHM model, however,

[1] Some even assert that a biased coin is a statistical unicorn—everyone talks about it but no one has actually encountered one [5]. Physics models support this assertion as long as the coin is not bent [12] or allowed to spin on the ground [3, 4].

mostly because it was not recorded whether the coin landed on the same side that it started. A notable exception is a sequence of 40,000 coin flips collected by Janet Larwood and Priscilla Ku [16]: Larwood always started the flips heads-up, and Ku always tails-up. Unfortunately, the results (i.e., 10,231/20,000 heads by Larwood and 10,014/20,000 tails by Ku) do not provide compelling evidence for or against the DHM hypothesis.

In order to carry out a diagnostic empirical test of the same-side bias hypothesized by DHM, we collected a total of 350,757 coin flips, a number that—to the best of our knowledge—dwarfs all previous efforts. To anticipate our main results, the data reveal overwhelming statistical evidence for the presence of same-side bias (and for individual differences in the extent of this bias). Furthermore, the data yield moderate evidence for the complete absence of a heads-tails bias. Exploratory analyses suggest the presence of a practice effect, as the degree of same-side bias decreases with the number of flips. The appendices demonstrate that the same conclusion obtains under a wide range of alternative analysis strategies.

Methods

Data collection

We collected data in three different settings using the same standardized protocol. First, a group of five bachelor students collected at least 15,000 coin flips each as a part of their bachelor thesis project, contributing 75,036 coin flips in total. Second, we organized a series of on-site "coin flipping marathons" where 35 people spent up to 12 hours coin-flipping (see e.g., [blinded for review] for a video recording of one of the events), contributing a total of 203,440 coin flips.[2] Third, we issued a call for collaboration via Twitter, which resulted in an additional seven people contributing a total of 72,281 coin flips.

We encouraged people to flip coins of various currencies and denomination to ascertain the generalizability of the effect. Furthermore, we encouraged coin tossers to exchange coins, as this potentially allows people-specific effects to be disentangled from coin-specific effects. Overall, a group of 48 people[3] (i.e., all but three of the co-authors) tossed coins of 44 different currencies $\times$ denominations and obtained a total number of 350,757 coin flips.

The protocol required that each person collects sequences of 100 consecutive coin flips.[4] In each sequence, people randomly (or according to an algorithm) selected a starting position (heads-up or tails-up) of the first coin flip, flipped the coin, caught it in their hand, recorded the landing position of the coin (heads-up or tails-up), and proceeded with flipping the coin starting from the same side it landed in the previous trial (we decided for this "autocorrelated" procedure as it simplified recording of the outcomes). In case the coin was not caught in hand, the flip was designated as a failure, and repeated from the same starting position. To simplify the recording and minimize coding errors, participants usually marked sides of the coins with permanent marker. To safeguard the integrity of the data collection effort, all participants videotaped and uploaded recordings of their coin flipping sequences.[5] See [blinded for review] for the data and video recordings.

Audit

We randomly sampled and audited 90 sequences of 100 coin flips each. We verified the existence of the video recordings (with occasionally missing video recordings due to file corruption or recording equipment malfunction) and attempted to re-code the outcome of individual coin tosses from the video recordings. We encountered video recordings of varying quality and detail which made one-to-one matching of the original coded sequences and the re-coded audited sequences highly challenging. However, assessing the degree of same-side bias on the original vs. the audited sequences revealed that the original sequences contained a highly similar degree of same-side bias. As such, it seems implausible that the original sequences were affected by coding bias in favor of the same-side hypothesis.

Statistical analysis

The preregistered and exploratory analyses outlined below are all Bayesian and use informed prior distributions. Appendix C provides a frequentist treatment. The Bayesian and frequentist results lead to the same qualitative conclusions.

[2] Including 2,700 coin flips collected by the first two authors on a separate occasion.

[3] One of the bachelor students collected data with a family member who is counted as a "coin-tosser" but who did not qualify for co-authorship.

[4] Some sequences slightly varied in length due to issues with keeping track of the number of flips.

[5] There are occasional missing recordings due to failures of recording apparatus/lost files.

Preregistered analysis: Same-side bias

Prior to data collection, we preregistered ([blinded for review]) an informed Bayesian binomial test with k same-side outcomes out of N trials,

$$k \sim \text{Binomial}(\beta, N),$$

assuming that the coin flips are independently and identically distributed across people and coins. We specified two competing hypotheses for the binomial success parameter β, where success denotes the coin landing on the same side it started from:

$$\begin{aligned} \text{No same-side bias, } \mathcal{H}_0 &: \beta = 0.5 \\ \text{DHM same-side bias, } \mathcal{H}_1 &: \beta \sim \text{Beta}(5100, 4900)_{[0.5,1]}. \end{aligned} \tag{1}$$

The highly informed $\text{Beta}(5100, 4900)_{[0.5,1]}$ prior distribution was meant to adequately represent the DHM hypothesis of the same-side bias [13]. Specifically: (a) the prior distribution is centered at the same-side bias of approximately 0.01 (the DHM point prediction of about 51%), (b) the prior distribution allows for reasonable variability around the mode—its standard deviation of approximately 0.005 corresponds to a 95% prior CI on the same-side bias that ranges from 0.00 to 0.02 (aligned with the majority of the estimates in 13; however, note that 13 also recorded a couple of outliers with a bias higher than 0.02); and (c) the prior distribution is truncation at 0.50, restricting the hypothesis to predict only the kind of bias stipulated by DHM (i.e., a same-side bias; a probability below 0.50 would represent an opposite-side bias, which would be anomalous according to the DHM model). Note that this informed hypothesis is motivated by the precise prediction of the DHM hypothesis—and as such allows a meaningful test. Appendix A, 'Prior Sensitivity Analysis', verifies the robustness of our conclusions to alternative specifications of the prior distribution.

The degree of evidence for the competing hypotheses is quantified by the Bayes factor [17, 18, 19],

$$\text{BF}_{10} = \frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)},$$

with $\text{BF}_{10} > 1$ indicating evidence for the DHM hypothesis and $\text{BF}_{10} < 1$ indicating evidence for the null hypothesis.

Exploratory analysis: Heads-tails bias

A similar analysis was conducted to test for heads-tails bias. Again an informed Bayesian binomial test was specified with h heads outcomes out of N trials,

$$h \sim \text{Binomial}(\alpha, N),$$

assuming that the coin flips are independently and identically distributed across people and coins. We defined two competing hypotheses in terms of the binomial success parameter α, where success denotes the coin landing heads:

$$\begin{aligned} \text{No heads-tails bias, } \mathcal{H}_{0a} &: \alpha = 0.5 \\ \text{Small heads-tails bias, } \mathcal{H}_{1a} &: \alpha \sim \text{Beta}(5000, 5000). \end{aligned} \tag{2}$$

In the absence of a theory-driven prediction for the size of the heads-tails bias, we used a prior distribution that is similar to the one we employed for the same-side bias, since: (a) if any heads-tails bias were present it would probably be minuscule—which is encoded in the 95% prior CI ranging from 0.49 (i.e., a 0.01 bias for tails) to 0.51 (i.e., a 0.01 bias for heads), and (b) we lack a reason to prefer either heads or tails, and hence the prior distribution is symmetric around $\alpha = 0.5$. Appendix A again verifies the robustness of our conclusions to alternative specifications of the prior distribution.

Exploratory analyses: Accounting for dependencies due to people and coins

In order to account for possible dependencies due to people and coins, the following hierarchical Bayesian logistic regression model was applied to the data. For the i^{th} flip of the k^{th} person with the j^{th} coin, the starting position $\text{y}_{t=0,ijk}$ and landing position $\text{y}_{t=1,ijk}$ (heads: $\text{y}_{t,ijk} = 1$, and tails: $\text{y}_{t,ijk} = 0$),

$$\begin{aligned} \mu_{ijk} &= \begin{cases} \text{logit}(\alpha_j) + \text{logit}(\beta_k) & \text{y}_{t=0,ijk} = 1 \\ \text{logit}(\alpha_j) - \text{logit}(\beta_k) & \text{y}_{t=0,ijk} = 0 \end{cases} \\ \text{y}_{t=1,ijk} &\sim \text{Bernoulli}(\text{logit}^{-1}(\mu_{ijk})), \end{aligned} \tag{3}$$

where logit(α_j) denotes the j^{th} coin's heads-tails bias (i.e., logit(α) $= 0$ corresponds to the absence of same-side bias, as the probability of the same side is 0.50) and logit(β_k) denotes the k^{th} person's same-side bias. The coin-specific and person-specific biases allow for differences from the overall heads-tails bias α_μ and the overall same-side bias β_μ through the addition of a coin-specific and person-specific random-effects γ_{α_j} and γ_{β_k} respectively:

$$\begin{aligned} \gamma_{\alpha_j} &\sim \text{Normal}(0, \sigma_\alpha^2) \\ \gamma_{\beta_k} &\sim \text{Normal}(0, \sigma_\beta^2) \\ \text{logit}(\alpha_j) &= \text{logit}(\alpha_\mu) + \gamma_{\alpha_j} \\ \text{logit}(\beta_k) &= \text{logit}(\beta_\mu) + \gamma_{\beta_k}. \end{aligned} \tag{4}$$

Since the DHM model specifies that the same-side bias results from the wobble introduced by the coin tosser (i.e., the person), the model does not feature coin-specific random-effects in the same-side bias (also see Appendix C for results suggesting the absence of between-coin heterogeneity in the same-side bias). Similarly, the model does not feature person-specific random-effects in the heads-tails bias as it appears implausible that some people would be more likely to flip heads than tails or visa versa. As detailed in the sections below, the model specified above will be used for both parameter estimation and hypothesis testing, but with different prior distributions [c.f., 18, 20].

Parameter estimation

For parameter estimation we specified slightly informed prior distributions, that is, prior distributions that assign the majority of mass to plausible values without imposing hard constraints that could prohibit learning in the case of model misspecification:

$$\begin{aligned} \alpha_\mu &\sim \text{Beta}(312, 312) \\ \beta_\mu &\sim \text{Beta}(312, 312) \\ \sigma_\alpha &\sim \text{Normal}_+(0, 0.04) \\ \sigma_\beta &\sim \text{Normal}_+(0, 0.04). \end{aligned} \tag{5}$$

For both the probability of heads α_μ and the probability of the same-side β_μ, we specified Beta$(312, 312)$ prior distributions. The Beta$(312, 312)$ prior distributions are centered at 0.50 (i.e., no bias) with a standard deviation of 0.02; we chose those values as the 95% prior CI contains biases up to 4 times larger than those predicted by the DHM theory which grants sufficient flexibility while ruling out implausibly large biases. For both the between-coin heterogeneity σ_α and between-people heterogeneity σ_β in the biases, we specified Normal$_+(0, 0.04)$ prior distributions. The Normal$_+(0, 0.04)$ prior distribution results in approximately half-Normal prior distributions with a standard deviation of 0.01 when transformed to probability scale; we chose those values as we believed that the between-coin and between-flipper heterogeneity in the biases would most likely be lower or equal to the bias predicted by DHM. Later visualizations confirmed that the specified prior distributions were not overly restrictive.

Hypothesis testing

The model from Equation 4 was used to estimate the same-side bias and the heads-tails bias while taking into account the dependency across people and coins. Furthermore, we used the model as a starting point for testing the hypotheses of the same-side bias, $\mathcal{H}_\beta$, and heads-tails bias, $\mathcal{H}_\alpha$, with the addition of hypotheses about between-people heterogeneity in the same-side bias, $\mathcal{H}_{\sigma_\theta}$, and between-coin heterogeneity in the heads-tails bias, $\mathcal{H}_{\sigma_\gamma}$:

$$\begin{aligned} &\mathcal{H}_{\beta,1} : \beta_\mu \sim \text{Beta}(5100, 4900)_{[0.5,1]} && \text{vs. } \mathcal{H}_{\beta_\mu,0} : \beta_\mu = 0.5 \\ &\mathcal{H}_{\alpha,1} : \alpha_\mu \sim \text{Beta}(5000, 5000) && \text{vs. } \mathcal{H}_{\alpha_\mu,0} : \alpha_\mu = 0.5 \\ &\mathcal{H}_{\sigma_\beta,1} : \sigma_\beta \sim \text{Gamma}(4, 200) && \text{vs. } \mathcal{H}_{\sigma_\beta,0} : \sigma_\beta = 0 \\ &\mathcal{H}_{\sigma_\alpha,1} : \sigma_\alpha \sim \text{Gamma}(4, 200) && \text{vs. } \mathcal{H}_{\sigma_\alpha,0} : \sigma_\theta = 0. \end{aligned} \tag{6}$$

In order to test the alternative hypotheses of the same-side bias $\mathcal{H}_{\beta,1}$ and the heads-tails bias $\mathcal{H}_{\alpha,1}$ we specified beta prior distributions that were identical to those used in the non-hierarchical models. For the between-people and between-coin heterogeneity in the same-side bias and in the heads-tails bias, we specified Gamma$(4, 200)$ prior distributions. This choice was motivated by the following considerations: (a) under the Gamma$(4, 200)$ priors, the expected heterogeneity of the biases (i.e., the expected standard deviation) equals 0.005 when transformed to the probability scale— half of the hypothesized effect; and (b) under the Gamma$(4, 200)$ priors, the standard deviation of the heterogeneity equals 0.0025 when transformed to the probability scale— half of the expected heterogeneity itself. Although these gamma prior distributions were not directly informed by the DHM theory, they are consistent with the estimates from DHM [13]; we consider these prior distributions reasonable and relevant for a hypothesis test. As before, Appendix A confirms the robustness of the conclusions under alternative prior specifications.

To test the hypotheses while accounting for uncertainty in the model structure, we used Bayesian model averaging [21, 22, 23] and specified 16 possible models as an orthogonal combination of the different possible hypotheses. For example, $\mathcal{M}_1$ specifies the presence of the same-side bias, the presence of the heads-tails bias, the presence of between-people heterogeneity in same-side bias, and the presence of between-people heterogeneity in the heads/tails bias. $\mathcal{M}_2$ specifies the presence of the same-side bias, the presence of the heads-tails bias, the presence of between-people heterogeneity in same-side bias, and the absence of between-people heterogeneity in the heads-tails bias. The last model, $\mathcal{M}_{16}$, then specifies the absence of the same-side bias, the absence of the heads-tails bias, the absence of between-people heterogeneity in same-side bias, and the absence of between-people heterogeneity in the heads/tails bias. The entire model space is listed as follows:

$$\begin{aligned}
\mathcal{M}_1 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_2 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_3 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_4 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_5 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_6 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_7 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_8 &= \mathcal{H}_{\beta,1}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_9 &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_{10} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_{11} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_{12} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,1}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_{13} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_{14} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,1}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}\\
\mathcal{M}_{15} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,1}\\
\mathcal{M}_{16} &= \mathcal{H}_{\beta,0}, \mathcal{H}_{\alpha,0}, \mathcal{H}_{\sigma_\beta,0}, \text{ and } \mathcal{H}_{\sigma_\alpha,0}.
\end{aligned}$$

Evidence for the parameters of interest may be quantified across the rival models using *inclusion Bayes factors*, a generalization of Bayes factors based on the change from prior to posterior odds [23]:

$$\underbrace{\text{BF}_{\text{AB}}}_{\substack{\text{Inclusion Bayes factor}\\ \text{for A vs. B}}} = \underbrace{\frac{\sum_{a\in A} p(\mathcal{M}_a \mid \text{y})}{\sum_{b\in B} p(\mathcal{M}_b \mid \text{y})}}_{\substack{\text{Posterior inclusion odds}\\ \text{for A vs. B}}} \Bigg/ \underbrace{\frac{\sum_{a\in A} p(\mathcal{M}_a)}{\sum_{b\in B} p(\mathcal{M}_b)}}_{\substack{\text{Prior inclusion odds}\\ \text{for A vs. B}}}, \tag{7}$$

where A contains a set of models where a given hypothesis holds and B contains the compliment. Specifically, a test for the presence vs. absence of same-side bias involves the comparison between $A \in \{1, 2, \ldots, 8\}$ vs. $B \in \{9, 10, \ldots, 16\}$; a test for the presence vs. absence of heads-tails bias involves the comparison between $A \in \{1, 2, 3, 4, 9, 10, 11, 12\}$ vs. $B \in \{5, 6, 7, 8, 13, 14, 15, 16\}$; a test for the between-people heterogeneity in the same-side bias involves the comparison between $A \in \{1, 2, 5, 6, 9, 10, 13, 14\}$ vs. $B \in \{3, 4, 7, 8, 11, 12, 15, 16\}$; finally, a test for the between-coin heterogeneity in the heads-tails bias involves the comparison between $A \in \{1, 3, 5, 7, 9, 11, 13, 15\}$ vs. $B \in \{2, 4, 6, 8, 10, 12, 14, 16\}$. Each of the 16 models was assigned an equal prior model probability, that is, $\Pr(\mathcal{M}_.) = {}^1/_{16}$

Exploratory analysis: Learning effects

We also explored the possibility of learning effects. Specifically, we wished to determine whether people's same-side bias $\text{logit}(\beta_k)$ (Equation 4) changes with an increasing number of tosses i. Learning curves may be specified in many ways [e.g., 24]. In an exploratory fashion, we visually examined the change in the observed proportion of same-side outcomes for each person with an increasing number of coin tosses and experimented with different models that extend Equation 3 in order to account for the learning effect. First we applied linear and quadratic effects of time but found that they underfitted the data (but see Appendix C for the frequentist analyses showing similar results). We also applied logistic functions with four and with five parameters [e.g., 25]; however, the random-effects versions of these models did not converge. We then ended up selecting a power function [26, 27]. Incorporating the power function into the hierarchical model (Equation 3) yields

$$\begin{aligned} \mu_{ijk} &= \begin{cases} \text{logit}(\alpha_j) + \Big(\text{logit}(\theta_k) + \text{logit}(\lambda_k)\, i_{jk}^{\rho_k}\Big) & \text{y}_{t=0,ijk} = 1 \\ \text{logit}(\alpha_j) - \Big(\text{logit}(\theta_k) + \text{logit}(\lambda_k)\, i_{jk}^{\rho_k}\Big) & \text{y}_{t=0,ijk} = 0 \end{cases} \\ \text{y}_{t=1,ijk} &\sim \text{Bernoulli}(\text{logit}^{-1}(\mu_{ijk})), \end{aligned} \tag{8}$$

which grants sufficient flexibility to accommodate a monotonic non-linear change in the same-side bias with an increasing number of tosses while retaining interpretability. The k^{th} person's same-side bias (previously $\text{logit}(\beta_k)$) is now decomposed into (1) the k^{th} person's baseline same-side bias $\text{logit}(\theta_k)$, and (2) the k^{th} person's toss-order dependent same-side bias $\text{logit}(\lambda_k)$ which, for each k^{th} person, either increases ($\rho_k > 0$) or decreases ($\rho_k < 0$) over time. If the k^{th} person's bias decreases over time, $\text{logit}(\theta_k)$ equals the asymptotic same-side bias. The k^{th} person's initial same-side bias corresponds to $\text{logit}(\beta_k) + \text{logit}(\lambda_k)$. The specification of the j^{th} coin's heads-tails bias $\text{logit}(\alpha_j)$ remains unchanged.

The hierarchical model specification is completed by adding random effects for the learning curve. In conjunction with the random-effects for the constant-bias model, this yields the following structure:

$$\begin{aligned} \gamma_{\alpha_j} &\sim \text{Normal}(0, \sigma_\alpha^2) \\ \gamma_{\theta_k} &\sim \text{Normal}(0, \sigma_\beta^2) \\ \gamma_{\rho_k} &\sim \text{Normal}(0, \sigma_\rho^2) \\ \gamma_{\lambda_k} &\sim \text{Normal}(0, \sigma_\lambda^2) \\ \text{logit}(\alpha_j) &= \text{logit}(\alpha_\mu) + \gamma_{\alpha_j} \\ \text{logit}(\beta_k) &= \text{logit}(\beta_\mu) + \gamma_{\beta_k} \\ \rho_k &= \rho_\mu + \gamma_{\rho_k} \\ \lambda_k &= \lambda_\mu + \gamma_{\lambda_k}, \end{aligned} \tag{9}$$

where $\text{logit}(\alpha_\mu)$ corresponds to the overall heads-tails bias, $\text{logit}(\theta_\mu)$ corresponds to the overall baseline same-side bias, $\text{logit}(\lambda_\mu)$ corresponds to the overall toss-order dependent same-side bias, and ρ_μ corresponds to the overall learning rate. The coin-specific and person-specific random-effects γ_{α_j}, γ_{θ_k}, γ_{λ_k}, and γ_ρ correspond to the coin-specific and person-specific deviations that are normally distributed with mean zero and the corresponding between-coin and between-person heterogeneities σ_α, σ_β, σ_ρ, and σ_λ.

The parameters of the learning effects model were estimated under the following prior distributions, which mimic those of the earlier hierarchical model without learning effects:

$$\begin{aligned} \alpha_\mu &\sim \text{Beta}(312, 312) \\ \theta_\mu &\sim \text{Beta}(312, 312) \\ \lambda_\mu &\sim \text{Beta}(312, 312) \\ \rho_\mu &\sim \text{Normal}(0, 10) \\ \sigma_\alpha &\sim \text{Normal}_+(0, 0.04) \\ \sigma_\theta &\sim \text{Normal}_+(0, 0.04) \\ \sigma_\lambda &\sim \text{Normal}_+(0, 0.04) \\ \sigma_\rho &\sim \text{Normal}_+(0, 1) \end{aligned} \tag{10}$$

The overall learning rate ρ_μ and the between-people heterogeneity in the learning rate σ_ρ were assigned $\text{Normal}(0, 10)$ and $\text{Normal}_+(0, 1)$ prior distributions, respectively, as these were sufficiently flexible to accommodate a wide range of learning rates.

In order to estimate the hierarchical learning rate model efficiently, the time series were aggregated into batches of about 100 coin flips performed with the same coin by the same person with the time ordinate i_{jk} averaged within the batch and scaled by 1000 for better interpretability. Post-analysis visualizations confirmed that the specified prior distributions were not overly restrictive.

All mixed-effect models were estimated in `Stan` [28] via the `rstan` R package [version 2.26.1 29]. The marginal likelihood was determined using bridge sampling [30, 31] via the `bridgesampling` R package [32, version 1.1-2]. The mixed-effect models were run using 10 chains with $15{,}000$ warm-up iterations and $10{,}000$ sampling iterations each. All models converged with $\hat{R} < 1.01$. Only the learning effects model showed divergent transitions. As mentioned before, Appendix A describes prior sensitivity analysis, Appendix B describes variability in the same-side bias analysis, and Appendix C reproduces our analyses in the frequentist framework.

Results

Descriptives

The data confirm the prediction from the DHM model: the coins landed how they started more often than 50%. Specifically, the data feature $178{,}079$ same-side landings from $350{,}757$ tosses, $\Pr(\text{same side}) = 0.5077$, 95% central credible interval (CI) $[0.5060, 0.5094]$ (under a uniform prior distribution, see Table 1 for a by-person summary), which is remarkably close to DHM's prediction of (approximately) 51%.

In addition, the data show no trace of a heads-tails bias. Specifically, we obtained $175{,}421$ heads out of $350{,}757$ tosses, $\Pr(\text{heads}) = 0.5001$, 95% CI $[0.4985, 0.5018]$ (under a uniform prior distribution, see Table 2 in the methods section for a by-coin summary).

Preregistered analysis: Same-side bias

The preregistered Bayesian informed binomial hypothesis test indicates extreme evidence in favor of the same-side bias predicted by the DHM model: $\text{BF}_{\text{same-side bias}} = 1.76 \times 10^{17}$ (see Appendix C for concordant frequentist result).

Exploratory analysis: Heads-tails bias

A similar (not-preregistered) analysis yields moderate evidence against the presence of a heads-tails bias, $\text{BF}_{\text{heads-tails bias}} = 0.168$ (see Appendix C for concordant frequentist result).

Exploratory analyses: Accounting for dependencies due to people and coins

With the data in hand we realized that the same-side bias was possibly subject to considerable between-people heterogeneity. Therefore we specified a more complex hierarchical Bayesian logistic regression model (Equation 3) that includes both heterogeneity in same-side bias between people and heterogeneity in heads-tails bias between coins; this hierarchical model was then used to estimate the parameters (Equation 5) and to test the hypotheses (Equation 6) using Bayesian model-averaging and inclusion Bayes factors (Equation 7). These analyses were not preregistered.

The posterior distribution of the probability of same side is slightly wider than in the preregistered nonhierarchical analysis, $\Pr(\text{same side}) = 0.5098$, 95% CI $[0.5050, 0.5147]$, which is caused by the substantial between-people heterogeneity in the probability of the coin landing on the same side, $\text{sd}_{\text{people}}(\Pr(\text{same side})) = 0.0156$, 95% CI $[0.0119, 0.0200]$. The additional uncertainty lowers the evidence for the same-side bias to $\text{BF}_{\text{same-side bias}} = 2359$, which however remains extreme (e.g., when the hypothesis of a same-side bias has a prior probability of 0.50, a Bayes factor of about 2359 results in a posterior probability of 0.9996). Consistent with the visual impression from Figure 1, the hierarchical model reveals overwhelming evidence for the presence of between-people heterogeneity in same-side bias, $\text{BF}_{\text{people heterogeneity}} = 3.10 \times 10^{24}$ (see Appendix C for concordant frequentist result; see Table 1 in Online Supplements for overview of the individual models).

Compared to the results from the nonhierarchical model, the posterior estimates of the overall probability of heads remains practically unchanged, $\Pr(\text{heads}) = 0.5005$, 95% CI $[0.4986, 0.5026]$ with virtually no between-coin heterogeneity, $\text{sd}_{\text{coins}}(\Pr(\text{heads})) = 0.0.0018$, 95% CI $[0.0001, 0.0047]$. The evidence against the presence of heads-tails bias also remains practically unchanged, $\text{BF}_{\text{heads-tails bias}} = 0.182$. The model shows moderate evidence against the presence of between-coin heterogeneity in heads-tails bias, $\text{BF}_{\text{coin heterogeneity}} = 0.178$ (see Appendix C for concordant frequentist result).

Exploratory analysis: Learning effects

Following up on a suggestion by a reviewer, a more extensive inspection of the data suggested that the degree of same-side bias was changing over time, necessitating an extension of the constant-bias model to accommodate power-law

Table 1: By-person summary of the probability of a same side landing.

Person	Same side	Flips	Coins	Proportion [95% CI]	Joined
XiaoyiL	780	1600	2	0.487 [0.463, 0.512]	Marathon-MSc
JoyceP	1126	2300	3	0.490 [0.469, 0.510]	Marathon-MSc
AndreeaZ	2204	4477	4	0.492 [0.478, 0.507]	Marathon-MSc
KaleemU	7056	14324	8	0.493 [0.484, 0.501]	Bc Thesis
FelipeFV	4957	10015	3	0.495 [0.485, 0.505]	Internet
ArneJ	1937	3900	4	0.497 [0.481, 0.512]	Marathon-MSc
AmirS	7458	15012	6	0.497 [0.489, 0.505]	Bc Thesis
ChrisGI	4971	10005	5	0.497 [0.487, 0.507]	Marathon-Manheim
FrederikA	5219	10500	5	0.497 [0.487, 0.507]	Internet
FranziskaN	5368	10757	3	0.499 [0.490, 0.508]	Internet
JasonN	3352	6700	7	0.500 [0.488, 0.512]	Marathon-PhD
RietvanB	1801	3600	4	0.500 [0.484, 0.517]	Marathon-PhD
PierreG	7506	15000	9	0.500 [0.492, 0.508]	Bc Thesis
KarolineH	2761	5500	5	0.502 [0.489, 0.515]	Marathon-PhD
SjoerdT	2510	5000	5	0.502 [0.488, 0.516]	Marathon-MSc
SaraS	5022	10000	3	0.502 [0.492, 0.512]	Marathon-Manheim
HenrikG	8649	17182	8	0.503 [0.496, 0.511]	Marathon
IrmaT	353	701	1	0.504 [0.467, 0.540]	Bc Thesis
KatharinaK	2220	4400	5	0.504 [0.490, 0.519]	Marathon-PhD
JillR	3261	6463	2	0.505 [0.492, 0.517]	Marathon
FrantisekB	10148	20100	11	0.505 [0.498, 0.512]	Marathon
IngeborgR	4340	8596	1	0.505 [0.494, 0.515]	Marathon
VincentO	2475	4900	5	0.505 [0.491, 0.519]	Marathon-MSc
EricJW	2071	4100	5	0.505 [0.490, 0.520]	Marathon-MSc
MalteZ	5559	11000	7	0.505 [0.496, 0.515]	Marathon-Manheim
TheresaL	1769	3500	4	0.505 [0.489, 0.522]	Marathon-MSc
DavidV	7586	14999	5	0.506 [0.498, 0.514]	Bc Thesis
AntonZ	5069	10004	2	0.507 [0.497, 0.516]	Marathon-Manheim
MagdaM	2510	4944	6	0.508 [0.494, 0.522]	Marathon-MSc
ThomasB	2540	5000	5	0.508 [0.494, 0.522]	Marathon-PhD
JonasP	5080	9996	5	0.508 [0.498, 0.518]	Marathon
BohanF	1118	2200	3	0.508 [0.487, 0.529]	Marathon-MSc
HannahA	1525	3000	4	0.508 [0.490, 0.526]	Marathon-MSc
AdrianK	1749	3400	3	0.514 [0.498, 0.531]	Marathon-MSc
AaronL	3815	7400	5	0.515 [0.504, 0.527]	Marathon-MSc
KoenD	3309	6400	7	0.517 [0.505, 0.529]	Marathon-PhD
MichelleD	2224	4300	5	0.517 [0.502, 0.532]	Marathon-PhD
RoyMM	2020	3900	4	0.518 [0.502, 0.534]	Marathon-MSc
TingP	1658	3200	4	0.518 [0.501, 0.535]	Marathon-MSc
MaraB	1426	2750	3	0.518 [0.500, 0.537]	Marathon-MSc
AdamF	4335	8328	2	0.520 [0.510, 0.531]	Marathon
AlexandraS	9080	17434	8	0.521 [0.513, 0.528]	Marathon
MadlenH	3705	7098	1	0.522 [0.510, 0.534]	Marathon
DavidKL	7895	15000	1	0.526 [0.518, 0.534]	Bc Thesis
XiaochangZ	1869	3481	4	0.537 [0.520, 0.553]	Marathon-MSc
FranziskaA	2055	3800	4	0.541 [0.525, 0.557]	Marathon-MSc
JanY	956	1691	2	0.565 [0.542, 0.589]	Marathon-MSc
TianqiP	1682	2800	3	0.601 [0.582, 0.619]	Marathon-MSc
Combined	178079	350757	44	0.508 [0.506, 0.509]	

Note. 'Proportion' refers to the observed proportion of coin flips that landed on the same side with a 95% central credible interval under uniform prior distributions (virtually identical to a frequentist confidence interval).

learning effects (i.e., Equation 8). In line with the exploratory nature of this modeling exercise, this hierarchical model was used to estimate parameters (cf. Equation 10), not to test hypotheses.

The top panel of Figure 3 reveals the probability of a same-side outcome is high initially but then decreases over the course of several thousands of coin flips; as the number of coin flips increases further, the same-side bias converges to a value near zero. The model suggests a large contribution of toss-order dependent same-side bias, $\Pr(\text{toss-order same side}) = 0.5250$, 95% CI [0.5129, 0.5374], which is accompanied with substantial between-people heterogeneity, $\text{sd}_{\text{people}}(\Pr(\text{toss-order same side})) = 0.0296$, 95% CI [0.0215, 0.0386]. The negative learning effect, $\rho = -1.6173$, 95% CI $[-2.8269, -0.8881]$, is also accompanied by substantial between-people heterogeneity,

Table 2: By-coin summary of the probability of heads.

Coin	Heads	Flips	People	Proportion [95% CI]
0.25CAD	48	100	1	0.480 [0.379, 0.582]
20DEM-silver	484	1000	1	0.484 [0.453, 0.515]
5CZK	1222	2500	2	0.489 [0.469, 0.509]
0.05NZD	984	2011	1	0.489 [0.467, 0.511]
0.10EUR	4515	9165	6	0.493 [0.482, 0.503]
1DEM	2464	5000	5	0.493 [0.479, 0.507]
50CZK	3207	6500	7	0.493 [0.481, 0.506]
2HRK	4258	8596	1	0.495 [0.485, 0.506]
1MXN	4180	8434	1	0.496 [0.485, 0.506]
1SGD	7655	15400	2	0.497 [0.489, 0.505]
5ZAR	3645	7325	1	0.498 [0.486, 0.509]
2EUR	24276	48772	28	0.498 [0.493, 0.502]
0.01GBP	498	1000	1	0.498 [0.467, 0.529]
0.50EUR	28617	57445	32	0.498 [0.494, 0.502]
0.20EUR	15665	31373	20	0.499 [0.494, 0.505]
0.25BRL	1998	4000	2	0.499 [0.484, 0.515]
0.10RON	1000	2001	1	0.500 [0.478, 0.522]
1CHF	2249	4500	4	0.500 [0.485, 0.514]
1EUR	18920	37829	25	0.500 [0.495, 0.505]
0.20GEL	4501	8998	5	0.500 [0.490, 0.511]
1CAD	5604	11200	11	0.500 [0.491, 0.510]
2CAD	1502	3000	3	0.501 [0.483, 0.519]
2MAD	1503	3000	1	0.501 [0.483, 0.519]
100JPY	752	1500	1	0.501 [0.476, 0.527]
2CHF	2259	4503	2	0.502 [0.487, 0.516]
5MAD	1007	2001	1	0.503 [0.481, 0.525]
0.20GBP	1516	3005	2	0.504 [0.486, 0.523]
1CNY	757	1500	1	0.505 [0.479, 0.530]
1CZK	505	1000	1	0.505 [0.474, 0.536]
2ILS	506	1000	1	0.506 [0.475, 0.537]
5JPY	1772	3500	2	0.506 [0.490, 0.523]
5SEK	8052	15902	7	0.506 [0.499, 0.514]
0.25USD	2180	4300	4	0.507 [0.492, 0.522]
1MAD	1014	2000	1	0.507 [0.485, 0.529]
0.50RON	1442	2844	3	0.507 [0.488, 0.526]
0.05EUR	3821	7514	6	0.509 [0.497, 0.520]
0.50GBP	765	1504	1	0.509 [0.483, 0.534]
2BDT	2038	4003	2	0.509 [0.494, 0.525]
10CZK	4572	8905	7	0.513 [0.503, 0.524]
0.20CHF	518	1000	1	0.518 [0.487, 0.549]
0.50SGD	1449	2781	3	0.521 [0.502, 0.540]
0.02EUR	158	300	1	0.527 [0.468, 0.584]
1GBP	791	1500	2	0.527 [0.502, 0.553]
2INR	552	1046	1	0.528 [0.497, 0.558]
Combined	175421	350757	48	0.500 [0.498, 0.502]

Note. ‘Proportion’ refers to the observed proportion of coin flips that landed heads with a 95% central credible interval under uniform prior distributions (virtually identical to a frequentist confidence interval).

$\sigma_\rho = 0.8406$, 95% CI $[0.3660, 1.6521]$. The negative learning effect suggests a decrease to a baseline same-side bias. The baseline same-side bias which serves as the asymptotic bias for the infinite coin toss is indistinguishable from chance, $\Pr(\text{baseline same side}) = 0.5014$, 95% CI $[0.4980, 0.5046]$, and barely differs between people, $\text{sd}_{\text{people}}(\Pr(\text{baseline same side})) = 0.0028$, 95% CI $[0.0001, 0.0067]$. The same-side bias estimate for the first flip corresponds to the sum of the toss-order dependent bias and the baseline bias, $\Pr(\text{initial same side}) = 0.5264$, 95% CI $[0.5151, 0.5385]$.

The left bottom panel of Figure 3 visually confirms that the posterior distribution for the asymptotic bias (solid line) is concentrated near chance, whereas the posterior distribution for the initial probability of a same-side outcome (dotted-dashed line) is markedly higher than chance. The posterior distributions for the associated heterogeneities are shown in the right bottom panel of Figure 3: the between-people variability in the asymptotic bias (solid line) is practically absent, whereas the between-people variability in the initial probability of a same-side outcome (dotted-dashed line) is relatively

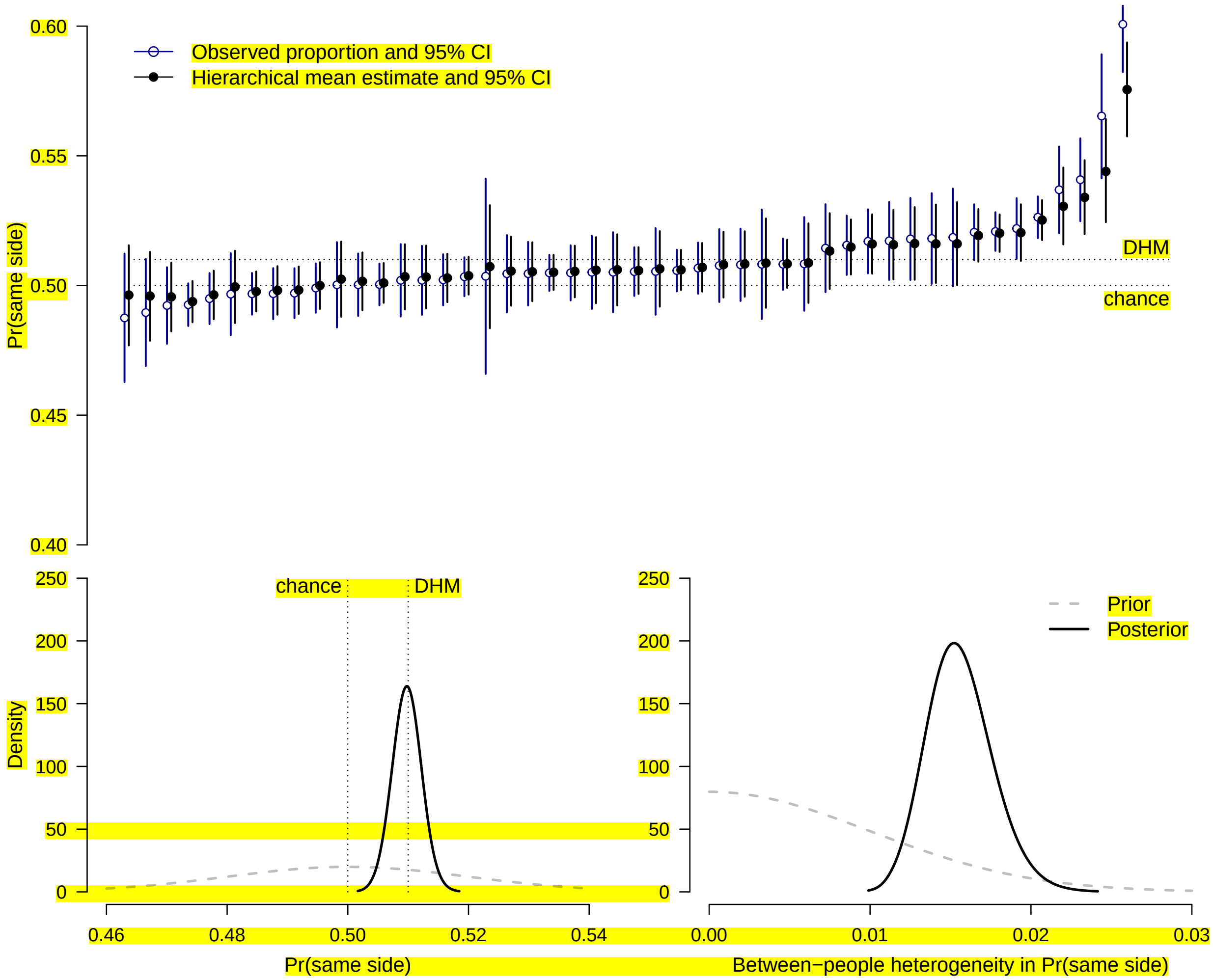

Figure 1: Coins have a tendency to land on the same side they started, confirming the predictions from the Diaconis, Holmes, and Montgomery (DHM) model of coin flipping. Top panel: posterior estimates of the probability of same side separately for each person, as obtained from the hierarchical Bayesian model with weakly informative, estimation-tailored prior distributions described in the methods section; Bottom-left panel: prior and posterior distributions for the overall probability of same side; Bottom-right panel: prior and posterior distributions for the between-people heterogeneity in the probability of the same side.

high. The attempt to estimate individual learning trajectories was frustrated by the fact that the person-level data are too variable to draw meaningful inferences; consequently, any particular model parameterization overwhelmingly pools the by-person trajectories towards the grand mean (see Appendix C for concordant frequentist result using linear and quadratic time effects and Online Supplementary Materials' Figure 6 showing a large heterogeneity in descriptive individual trajectories).

The posterior estimates of the overall probability of heads remains practically unchanged, $\text{Pr(heads)} = 0.5005$, 95% CI [0.4986, 0.5026] with virtually no between-coin heterogeneity, $\text{sd}_{\text{coins}}(\text{Pr(heads)}) = 0.0018$, 95% CI [0.0001, 0.0047].

Outlier exclusion

We repeated the statistical analyses after excluding four potential outliers with same-side sample proportions larger than 53% (i.e., the four largest and right-most estimates in the top panel of Figure 1). In general, the exclusion did not qualitatively affect our conclusions, although –as may be expected–the same-size bias decreased in size, and the between-people heterogeneity became less pronounced (see Appendix C for concordant frequentist result). Each of the four participants with the same-side sample proportion larger than 53% contributed fewer than 4,000 coin flips,

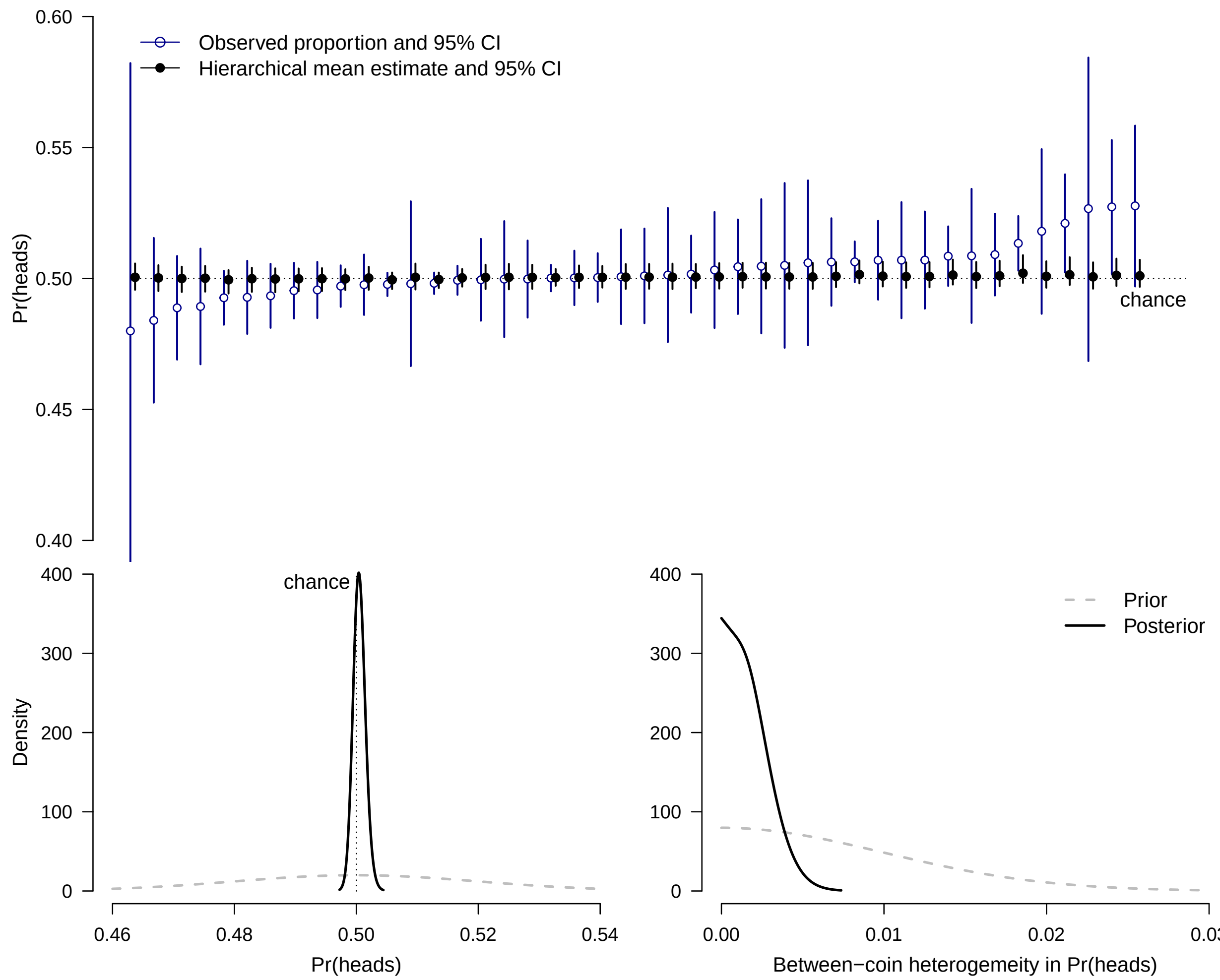

Figure 2: Coins have a tendency to land on heads and tails with equal probability, supporting the predictions from the Standard model of coin flipping. Top panel: posterior estimates of the probability of heads separately for each coin, as obtained from the hierarchical Bayesian model with weakly informative, estimation-tailored prior distributions described in the methods section; Bottom-left panel: prior and posterior distributions for the overall probability of heads; Bottom-right panel: prior and posterior distributions for the between coin heterogeneity in the probability of heads.

which raises a possibility that the observed proportion is a result of the a) initially larger same-side bias (before the learning effects occurs) and b) sampling variability. See additional robustness checks in Tables 3, 4, and 5 in the Online Supplements showing the qualitative conclusions do not change when excluding 1 to 5 participants with the lowest and the largest proportion of same-side outcomes.

After excluding the four potential outliers, the data feature 171,517 same-side landings from 338,985 tosses, $\text{Pr}(\text{same side}) = 0.5060$, 95% CI $[0.5043, 0.5077]$. The evidence in favor of the DHM hypothesis using the preregistered Bayesian informed binomial hypothesis test decreased notably but remains extreme, $\text{BF}_{\text{same-side bias}} = 1.28 \times 10^8$. The proportion of heads remained practically identical, 169,635 heads out of 338,985 tosses, $\text{Pr}(\text{heads}) = 0.5004$, 95% CI $[0.4987, 0.5021]$, as did the moderate evidence against heads-tails bias $\text{BF}_{\text{heads-tails bias}} = 0.190$.

The exclusion of potential outliers similarly affects the inference from the hierarchical model: although the same-side bias decreases and the associated heterogeneity is reduced (i.e., $\text{Pr}(\text{same side}) = 0.5060$, 95% CI $[0.5031, 0.5089]$ and $\text{sd}_{\text{people}}(\text{Pr}(\text{same side})) = 0.0072$, 95% CI $[0.0050, 0.0099]$), the evidence for the presence of the same-side bias and the associated between-people heterogeneity remains extreme (i.e., $\text{BF}_{\text{same-side bias}} = 787$ and $\text{BF}_{\text{people heterogeneity}} = 2.87 \times 10^7$). For the probability of heads, the inference remains practically unchanged, both with respect to the size of the effect

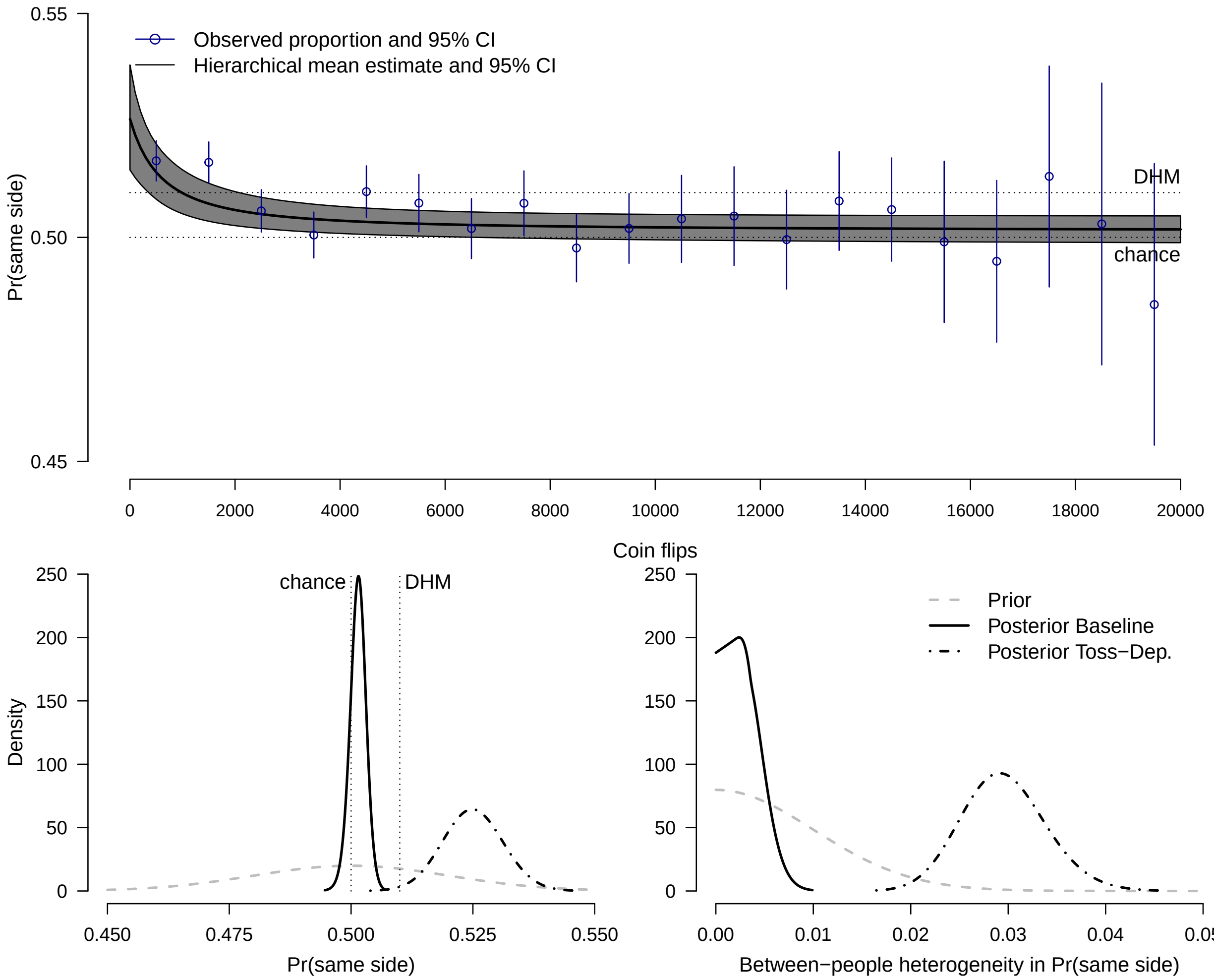

Figure 3: The degree of the individual same-side bias decreases as people flip more coins. Top panel: average sample proportions of same-side outcomes with 95% credible intervals (under uniform prior distributions), together with a summary of the posterior distribution for the group mean (under a hierarchical Bayesian model with weakly informative, estimation-tailored prior distributions as outlined in the main text), as a function of an increasing number of coin flips shown in batches of $1{,}000$ observations. Bottom-left panel: prior (grey) and posterior distributions for the baseline (black solid line) and toss-order dependent (dotted-dashed line) probability of a same-side outcome; Bottom-right panel: prior and posterior distributions for the between-people heterogeneity in the baseline and toss-order dependent same-side bias.

and associated heterogeneity (i.e., $\text{Pr(heads)} = 0.5008$, 95% CI $[0.4988, 0.5030]$ and $\text{sd}_{\text{coins}}(\text{Pr(heads)}) = 0.0020$, 95% CI $[0.0001, 0.0050]$), and with respect to the evidence for the presence of the heads-tails bias and the associated between-coin heterogeneity (i.e., $\text{BF}_{\text{heads-tails bias}} = 0.213$ and $\text{BF}_{\text{coin heterogeneity}} = 0.221$; see Table 2 in the Online Supplements for an overview of the individual models).

The exclusion of potential outliers affected the learning model in a similar manner. The contribution of the toss-order dependent same-side bias to the same-side bias slightly decreased, $\text{Pr(toss-order same side)} = 0.5177$, 95% CI $[0.5065, 0.5292]$, and the associated between-people heterogeneity was reduced, $\text{sd}_{\text{people}}(\text{Pr(toss-order same side)}) = 0.0218$, 95% CI $[0.0135, 0.0312]$. The negative learning effect decreased, $\rho = -1.4614$, 95% CI $[-3.2027, -0.6241]$, as did the associated between-people heterogeneity, $\sigma_\rho = 0.5587$, 95% CI $[0.0309, 1.6081]$. The baseline same-side bias remains indistinguishable from chance, $\text{Pr(baseline same side)} = 0.5011$, 95% CI $[0.4966, 0.5045]$, and barely differs between people, $\text{sd}_{\text{people}}(\text{Pr(baseline same side)}) = 0.0028$, 95% CI $[0.0001, 0.0067]$. The same-side bias estimate for the first flip also decreased, $\text{Pr(initial same side)} = 0.5188$, 95% CI $[0.5089, 0.5300]$. The overall probability of heads and the between-coin heterogeneity in the probability of heads, $\text{Pr(heads)} = 0.5008$, 95% CI $[0.4989, 0.5030]$

and $\mathrm{sd}_{\mathrm{coins}}(\Pr(\mathrm{heads})) = 0.0020$, 95% CI [0.0001, 0.0050] remains essentially unchanged. The shape of the learning trajectory also remains similar, and is still characterized by a steep decrease in the first couple of thousand coin flips (see Figure 3 in the Online Supplementary Materials).

Discussion

We collected 350,757 coin flips and found strong empirical confirmation of the counterintuitive and precise prediction from DHM model of human coin tossing: when people flip a coin, it tends to land on the same side as it started. Moreover, the data revealed a substantial degree of between-people variability in same-side bias: as can be seen from Figure 1, some people appear to have little or no same-side bias, whereas others do display a same-side bias, albeit to a varying degree. This variability is consistent with DHM model, in which the same-side bias originates from off-axis rotations (i.e., precession or wobbliness), which can reasonably be assumed to vary between people. Additional exploratory analyses suggest that the degree of the same-side bias decreases with the number of coin flips. A possible explanation of this decreasing bias is a coin-tossing practice effect—the more coins people flip, the closer they approach the 'perfect' wobble-less flip. Our results suggest that around 10,000 coin flips ($\approx$ 10 hours of coin flipping) might be enough to virtually eliminate the same-side bias.

Our results are aligned with previous empirical data evaluating the same-side bias; the DHM [13] 27 high-speed camera flips that initially suggested the same-side bias of approximately 1%, and the subsequent 20,000 coin flips by Janet Larwood and Priscilla Ku who found 1.2% and 0.1% same-side bias, respectively [16]. Future work may attempt to verify whether 'wobbly tossers' show a more pronounced same-side bias than 'stable tossers' and examine the learning effects in more detail. Furthermore, the present learning effects suggest that the same-side bias is best studied through many people who each contribute only a few thousand coin flips rather than through few people who each contribute tens of thousands of coin flips; the same-side bias is much more pronounced at the beginning of the experiment, and hence there are diminishing returns when a single person contributes very many coin flips. However, the effort required to test the more detailed hypotheses appears to be excessive, as this would ideally involve detailed analyses of high-speed camera recordings for individual flips (cf. 13).

In order to ensure the quality of the data, we videotaped and audited the data collection procedure (see the method section for details). The audit did not reveal anything suspicious (i.e., all participants performed the coin flips they reported); however, the video recordings were of insufficient quality to provide additional insights into the variability of the same-side bias. There also remains a legitimate concern: at the time when people were flipping the coins they were aware of the main hypothesis under test. Therefore it cannot be excluded that some of the participants were able to manipulate the coin flip outcomes in order to produce the same-side bias. In light of the nature of the coin tossing process, the evidence from the video recordings, and the precise correspondence between the data and the predictions from DHM model, we deem this possibility as unlikely; future work is needed to disprove it conclusively (e.g., by concealing the aim of the study).

Could future coin tossers use the same-side bias to their advantage? The magnitude of the observed bias can be illustrated using a betting scenario. If you bet a dollar on the outcome of a coin toss (i.e., paying 1 dollar to enter, and winning either 0 or 2 dollars depending on the outcome) and repeat the bet 1,000 times, knowing the starting position of the coin toss would earn you 19 dollars on average. This is more than the casino advantage for 6 deck blackjack against an optimal-strategy player, where the casino would make 5 dollars on a comparable bet, but less than the casino advantage for single-zero roulette, where the casino would make 27 dollars on average [33]. These considerations lead us to suggest that when coin flips are used for high-stakes decision-making, the starting position of the coin is best concealed.

Acknowledgments

Funding

The authors have no funding to declare, and conducted this research in their spare time.

Competing interests

The authors have no competing interest to declare.

Data and materials availability

All data and materials are available at `https://osf.io/pxu6r/`.

Ethical approval

The research project was approved by the Ethics Review Board of the Faculty of Social and Behavioral Sciences, University of Amsterdam, The Netherlands (2022-PML-15687).

Author note

During a revision we corrected two minor errors in the data set: (a) two coins were miscoded as different currency $\times$ denomination, and (2) one invalid coin flip with unknown starting position was not removed from the analysis as incorrectly treated as starting heads and landing tails. We fixed the coin coding issue, which reduced the number of unique coins from 46 to 44. We fixed the mistake in data processing and asked the participant to provide one additional valid coin flip to retain the same number of total coin flips. The additional flip started tails and landed tails, which increased the number of coins landing on the same side by one and decreased the number of coins landing on heads by one (we could not retrieve the original coin; consequently, the number of flips with 5 ZAR reduced by one and number of flips with 0.05 EUR increased by one). Our original findings were qualitatively unaffected by these errors.

References

[1] Kerrich JE. *An experimental introduction to the theory of probability*. E. Munksgaard . 1946.

[2] Feller W. *An introduction to probability: Theory and examples*. Wiley . 1957.

[3] Jaynes ET. *Probability theory: The logic of science*. Cambridge University Press . 2003.

[4] Küchenhoff H. Coin tossing and spinning–useful classroom experiments for teaching statistics. In: Physica-Verlag HD. 2008 (pp. 417–426).

[5] Gelman A, Nolan D. You can load a die, but you can't bias a coin. *The American Statistician* 2002; 56(4): 308–311.

[6] Macrobius AAT. Saturnalia. ≈ 431AD.

[7] Vulović VZ, Prange RE. Randomness of a true coin toss. *Physical Review A* 1986; 33(1): 576–582.

[8] Keller JB. The probability of heads. *The American Mathematical Monthly* 1986; 93(3): 191–197.

[9] Engel E. *A road to randomness in physical systems*. 71. Springer-Verlag . 1992.

[10] Strzałko J, Grabski J, Stefański A, Perlikowski P, Kapitaniak T. Dynamics of coin tossing is predictable. *Physics Reports* 2008; 469(2): 59–92.

[11] Strzałko J, Grabski J, Stefański A, Perlikowski P, Kapitaniak T. Understanding coin-tossing. *The Mathematical Intelligencer* 2010; 32: 54–58.

[12] Woo DK, Oh CH. Bent coin toss probability. *Journal of the Korean Data and Information Science Society* 2010; 21(1): 147–153.

[13] Diaconis P, Holmes S, Montgomery R. Dynamical bias in the coin toss. *SIAM Review* 2007; 49(2): 211–235.

[14] Buffon GL. Essai d'arithmétique morale. *Supplément à l'Histoire Naturelle* 1777; 4: article XXIII.

[15] Pearson K. *The chances of death, and other studies in evolution*. Edward Arnold . 1897.

[16] Berkley . 40,000 coin tosses yield ambiguous evidence for dynamical bias. `https://www.stat.berkeley.edu/~aldous/Real-World/coin_tosses.html`; 2009.

[17] Jeffreys H. Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophy Society* 1935; 31: 203–222.

[18] Jeffreys H. *Theory of probability*. Oxford, UK: Oxford University Press. 1 ed. 1939.

[19] Kass RE, Raftery AE. Bayes factors. *Journal of the American Statistical Association* 1995; 90(430): 773–795.

[20] Jeffreys . *Theory of probability*. Oxford University Press. 3rd edition ed. 1961.

[21] Raftery AE, Madigan D, Volinsky CT. Accounting for model uncertainty in survival analysis improves predictive performance. In: University Press; 1995: 323–349.

[22] Hoeting JA, Madigan D, Raftery AE, Volinsky CT. Bayesian model averaging: A tutorial. *Statistical Science* 1999; 14(4): 382–401.

[23] Hinne M, Gronau QF, van den Bergh D, Wagenmakers EJ. A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science* 2020; 3(2): 200–215.

[24] Viering T, Loog M. The shape of learning curves: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2022; 45(6): 7799–7819.

[25] Gottschalk PG, Dunn JR. The five-parameter logistic: a characterization and comparison with the four-parameter logistic. *Analytical Biochemistry* 2005; 343(1): 54–65.

[26] Box GE, Tidwell PW. Transformation of the independent variables. *Technometrics* 1962; 4(4): 531–550.

[27] Logan GD. The Weibull distribution, the power law, and the instance theory of automaticity. *Psychological Review* 1995; 102: 751–756.

[28] Carpenter B, Gelman A, Hoffman MD, et al. Stan: A probabilistic programming language. *Journal of Statistical Software* 2017; 76(1): 1–32.

[29] Stan Development Team . RStan: the R interface to Stan. http://mc-stan.org/; 2018. R package version 2.17.3.

[30] Meng XL, Wong WH. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica* 1996: 831–860.

[31] Gronau QF, Sarafoglou A, Matzke D, et al. A tutorial on bridge sampling. *Journal of Mathematical Psychology* 2017; 81: 80–97.

[32] Gronau QF, Singmann H, Wagenmakers EJ. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software* 2020; 92(10): 1–29.

[33] Hannum RC. A guide to casino mathematics. *UNLV Gaming Studies Research Center* 2003.

[34] Johnson VE, Rossell D. On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2010; 72(2): 143–170.

[35] Johnson VE, Pramanik S, Shudde R. Bayes factor functions for reporting outcomes of hypothesis tests. *Proceedings of the National Academy of Sciences* 2023; 120(8): e2217331120.

[36] Pawel S. A Bayes factor framework for unified parameter estimation and hypothesis testing. *arXiv preprint* 2024.

[37] Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 2015; 67(1): 1–48.

[38] White T, Noble D, Senior A, Hamilton WK, Viechtbauer W. metadat: Meta-Analysis Datasets. 2022. R package version 1.3-0.

[39] Viechtbauer W. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* 2010; 36(3): 1–48.

Appendix A: Prior sensitivity analysis

We assessed the sensitivity of the performed hypothesis tests to the prior specification. First, we examined prior sensitivity of the Bayesian (nonhierarchical) informed binomial hypothesis tests; second, we examined prior sensitivity of the more complex hierarchical Bayesian logistic regression utilizing Bayesian model averaging. We specified a range of prior distributions for each alternative hypothesis by employing the normal-moment prior distribution [34], Normal-Moment(ϕ),

$$p(x \mid \phi) = \frac{2x^2}{\sqrt{\pi}|\phi^3|}\exp\left(\frac{-x^2}{\phi^2}\right),$$

which allows us to specify non-local prior distributions using a single parameter—the mode of the prior distribution ϕ. When the support of the Normal-Moment distribution is unrestricted, it features two modes at $\pm\phi$ with the density decreasing towards zero at 0 and towards zero at $\pm\infty$. Varying the ϕ parameter across a range of plausible values allows us to assess the evidence in favor or against the hypothesis as a function of the prior distribution's mode, sometimes described as the Bayes factor function [BFF, 35]. We examined the prior sensitivity with both the full data set and on data set after excluding potential outliers (i.e., people with the observed probability of the same side larger than 0.53).

All prior sensitivity analyses were conducted using `Stan` [28] via the `rstan` `R` package [version 2.26.1 29] and estimated the marginal likelihood using bridge sampling [30, 31] using the `bridgesampling` `R` package [32, version 1.1-2]. The models were run using 4 chains with 15,000 warm-up and 10,000 sampling iterations.

For an alternative approach (an inversion of Bayes factor function) see Section 4.1 and 4.2 in [36], which leads to similar conclusions.

Nonhierarchical analyses

Using the Normal-Moment prior distributions results in the modified definition of our hypotheses about the same-side bias (Equation 1):

$$\begin{aligned} \text{No same-side bias, } \mathcal{H}_0 &: \text{logit}(\beta) = 0.5 \\ \text{DHM same-side bias, } \mathcal{H}_1 &: \text{logit}(\beta) \sim \text{Normal-Moment}_+(\phi_\beta), \end{aligned}$$

and the heads-tails bias (Equation 2):

$$\begin{aligned} \text{No heads-tails bias, } \mathcal{H}_{0a} &: \text{logit}(\alpha) = 0.5 \\ \text{Small heads-tails bias, } \mathcal{H}_{1a} &: \text{logit}(\alpha) \sim \text{Normal-Moment}(\phi_\alpha). \end{aligned}$$

The main difference lies in defining prior distributions for the same-side bias logit(β) and heads-tails bias logit(α) instead of the probability of same side β and probability of heads α.

We examined a range of modes $\phi_\beta = [0, 0.08]$ and $\phi_\alpha = [0, 0.08]$ of the Normal-Moment distributions. The examined modes translate to the mode probability of the same side ranging from 0.50 to 0.52 and the probability of heads ranging from 0.48 to 0.52. The upper ranges of the prior modes were chosen a posteriori based on the observed ranges of estimates

Results

The left panel of Figure 4 visualizes the Bayes factor functions for the same-side bias using the complete data set (full line) and after excluding the potential outliers (dashed line). The evidence in favor of the same-side bias is rapidly increasing even for small deviations from the null hypothesis; we find strong evidence ($\text{BF}_{\text{same-side bias}} > 10$) for all examined modes of alternative hypothesis of the presence of the same-side bias (i.e., Pr(same side > 0.5003)) regardless of potential outlier exclusion. The alternative hypothesis of the same-side bias receives the highest support at the mode Pr(same side $= 0.5062$) ($\text{BF}_{\text{same-side bias}} = 2.85 \times 10^{17}$) using the full data set and at the mode Pr(same side $= 0.5047$) ($\text{BF}_{\text{same-side bias}} = 1.01 \times 10^{10}$) after excluding potential outliers.

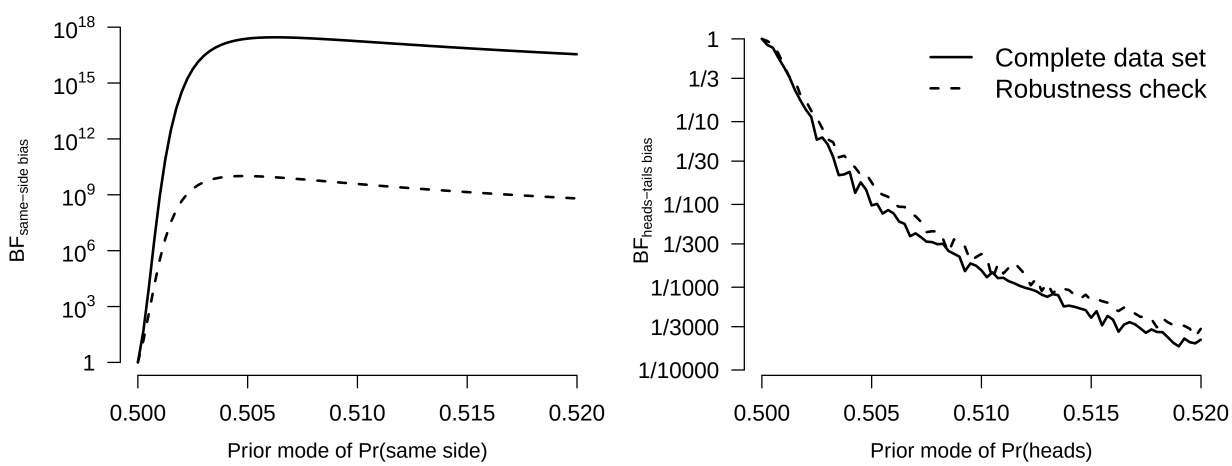

Figure 4: A prior sensitivity analysis demonstrates robustness of the same-side bias (left) and the absence of the heads-tails bias (right) across different prior specifications (x-axis) and exclusion of potential outliers (full line vs. dashed line) from nonhierarchical models.

The right panel of Figure 4 visualizes the Bayes factor functions for the heads-tails bias using the complete data set (full line) and after excluding the potential outliers (dashed line). The evidence against the heads-tails bias is increasing for any deviations from the null hypothesis; we find strong evidence against the heads-tails bias ($\text{BF}_{\text{heads-tails bias}} < 1/10$) for all examined modes of alternative hypothesis of the heads-tails bias starting at Pr(heads > 0.5025) using the full data set and at Pr(same side > 0.5028) after excluding potential outliers.

Analyses accounting for dependencies due to people and coins

Using the Normal-Moment prior distributions results in the modified definition of our hypotheses (Equation 6):

$$\begin{aligned}
&\mathcal{H}_{\beta,1} : \text{logit}(\beta) \sim \text{Normal-Moment}_{+}(\phi_\beta) && \text{vs. } \mathcal{H}_{\beta,0} : \beta = 0.5 \\
&\mathcal{H}_{\alpha,1} : \text{logit}(\alpha) \sim \text{Normal-Moment}(\phi_\alpha) && \text{vs. } \mathcal{H}_{\alpha,0} : \alpha = 0.5 \\
&\mathcal{H}_{\sigma_\theta,1} : \sigma_\theta \sim \text{Normal-Moment}_{+}(\phi_{\sigma_\theta}) && \text{vs. } \mathcal{H}_{\sigma_\theta,0} : \sigma_\theta = 0 \\
&\mathcal{H}_{\sigma_\gamma,1} : \sigma_\gamma \sim \text{Normal-Moment}_{+}(\phi_{\sigma_\gamma}) && \text{vs. } \mathcal{H}_{\sigma_\gamma,0} : \sigma_\gamma = 0.
\end{aligned} \tag{11}$$

We examined the sensitivity of each hypothesis test marginally to the non-assessed hypotheses (i.e., we fixed the prior mode of all but the tested parameters) to keep the computational load feasible. This corresponds to computing four BFFs under the corresponding settings for the alternative hypotheses (the null hypotheses remain unchanged):

$$\begin{aligned}
&\text{BFF}_\beta : \phi_\beta = [0, 0.08], \phi_\alpha = 0.04, \phi_{\sigma_\theta} = 0.02, \text{and } \phi_{\sigma_\gamma} = 0.02 \\
&\text{BFF}_\alpha : \phi_\beta = 0.04, \phi_\alpha = [0, 0.08], \phi_{\sigma_\theta} = 0.02, \text{and } \phi_{\sigma_\gamma} = 0.02 \\
&\text{BFF}_\beta : \phi_\beta = 0.04, \phi_\alpha = 0.04, \phi_{\sigma_\theta} = [0, 0.08], \text{and } \phi_{\sigma_\gamma} = 0.02 \\
&\text{BFF}_\beta : \phi_\beta = 0.04, \phi_\alpha = 0.04, \phi_{\sigma_\theta} = 0.02, \text{and } \phi_{\sigma_\gamma} = [0, 0.08].
\end{aligned}$$

The fixed prior modes were chosen to correspond approximately to the mean of prior distributions specified in Equation 6, that is, a same-side bias $\beta \approx 0.51$, heads-tails bias $\alpha \approx 0.49$ and 0.51, between-people heterogeneity in the same-side bias $\sigma_\theta \approx 0.01$, and between-coin heterogeneity in the heads-tails bias $\sigma_\theta \approx 0.01$. The upper ranges of the prior modes were chosen a posteriori, based on the observed ranges of estimates, i.e., up to a bias of ≈ 0.52 or between-people / coin heterogeneity in the biases of ≈ 0.02.

Results

The first row of Figure 5 visualizes the Bayes factor functions for the same-side bias (left) and heads-tails bias (right) when accounting for between people and between coin dependencies using the complete data set (full line) and after excluding the potential outliers (dashed line). In contrast to the dependency unadjusted analysis (cf., Figure 4), the degree of evidence in favor of the same-side bias and against heads-tails bias is notably smaller. Nevertheless, there is strong evidence ($\text{BF}_{\text{same-side bias}} > 10$) for all examined modes of alternative hypothesis of the presence of the same-side bias (i.e., $\text{Pr}(\text{same side} > 0.5013)$) regardless of potential outlier exclusion. The alternative hypothesis of the same-side bias receives the highest support at the mode $\text{Pr}(\text{same side} = 0.5075)$ ($\text{BF}_{\text{same-side bias}} = 7208$) using the full data set and at the mode $\text{Pr}(\text{same side} = 0.5050)$ ($\text{BF}_{\text{same-side bias}} = 1697$) after excluding potential outliers. The heads-tails bias is decreasing slower than in the dependency unadjusted analysis and we find strong evidence for the absence of heads-tails bias ($\text{BF}_{\text{heads-tails bias}} < 1/10$) only against alternative hypotheses with mode $\text{Pr}(\text{heads} = 0.5037)$ or larger regardless of potential outlier exclusion.

The second row of Figure 5 visualizes the Bayes factor functions for the between-person heterogeneity in the same-side bias (left) and the between-coin heterogeneity in heads-tails bias (right) using the complete data set (full line) and after excluding the four potential outliers (dashed line). The degree of evidence in favor of the between-person heterogeneity in the same-side bias is extreme: we find strong evidence for the between-person heterogeneity in the same-side bias in all examined modes of alternative hypothesis, with the strongest evidence for the alternative hypothesis of the between-person heterogeneity at the mode $\text{sd}_{\text{people}}(\text{Pr}(\text{same side})) = 0.0137$ ($\text{BF}_{\text{between-people heterogeneity}} = 1.27 \times 10^{26}$) using the full data set and at the mode $\text{sd}_{\text{people}}(\text{Pr}(\text{same side})) = 0.0062$ ($\text{BF}_{\text{between-people heterogeneity}} = 4.15 \times 10^{7}$) after excluding potential outliers. The evidence for the absence of between-coin heterogeneity in heads-tails bias accumulates more slowly: we find strong evidence for the absence of between-coin heterogeneity in heads-tails bias ($\text{BF}_{\text{between-coin heterogeneity}} < 1/10$) only against alternative hypotheses with mode $\text{sd}_{\text{coins}}(p_{\text{heads}}) = 0.0050$ or larger regardless of potential outlier exclusion.

Summary

The prior sensitivity analysis verified that our findings are robust to the specific choices of our prior distributions. In fact, the prior sensitivity analysis suggests that the data support a substantive range of parameterization of the same-side bias and between-people heterogeneity in the same-side bias.

Appendix B: Same-side bias variability by recruitment site

We assessed whether the degree of the same-side bias systematically varies according to the source of participant's recruitment. We extended the same-side bias part of the Equation 3, $\text{logit}(\beta)_k$, with a by-recruitment site difference

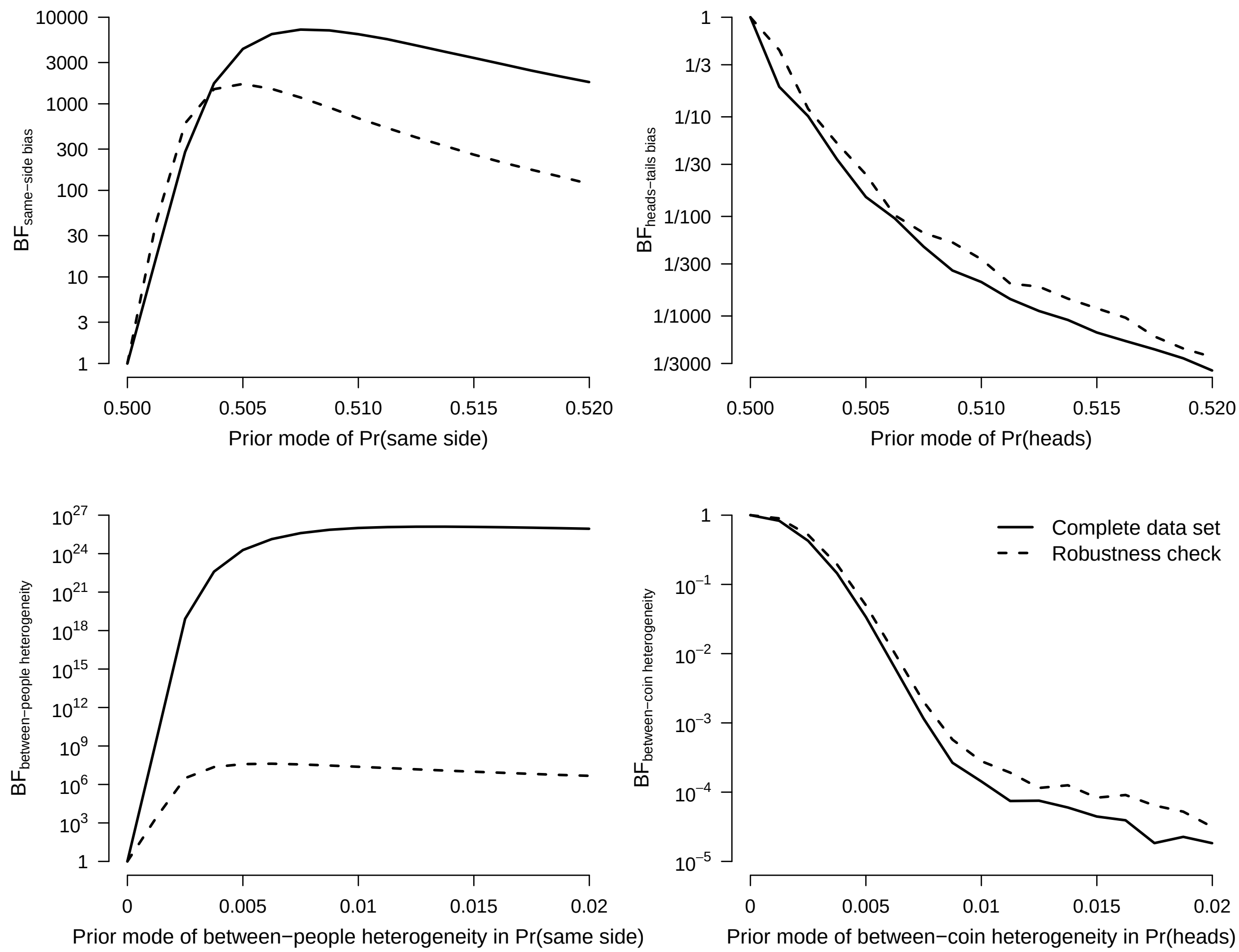

Figure 5: Prior sensitivity analysis demonstrates robustness of the same-side bias (top left), the absence of the heads-tails bias (top right), the presence of between-person heterogeneity in the same-side bias (bottom left), and the absence of between-coin heterogeneity in the heads-tails bias (bottom left) across different prior specifications (x-axis) and exclusion of the potential outliers (full vs. dashed line) from hierarchical models accounting for dependencies between people and coins.

from the same-side bias δ_k,

$$\text{logit}(\beta_k) = \text{logit}(\beta_\mu) + \gamma_{\beta_k} + \delta_k,$$

such that the δ_k implement a sum to zero constrain via standardized orthonormal contrast.

We only attempted to estimate the parameters and specified slightly informed prior distributions as in Equation 5 for α_μ, β_μ, σ_α, and σ_β and a slightly informed prior distribution for the by-recruitment site difference from the same-side bias

$$\delta_k \sim \text{Normal}(0, 0.20),$$

which results in standard deviation of 0.05 on probability scale. We chose this prior distribution over our originally intended prior distribution $\text{Normal}(0, 0.04)$ (i.e., leading to a 0.01 standard deviation on the probability scale), because the originally intended distribution turned out to restrict the posterior distribution too severely (the data contained very little information about the by-recruitment site differences).

Results

Figure 6 visualizes the prior and posterior distributions of by-recruitment site differences from the probability of the same side. Note that all posterior distributions remain relatively wide, highlighting a lack of information about potential person-level moderators. Out of all six sites, only the 95% CI for the 'Marathon-Msc' does not contain zero, $\delta_{\text{Marathon-Msc}} = 0.0010$, 95% CI $[0.0014, 0.0184]$, however, the difference disappears after excluding outliers. The remaining by-recruitment site differences: $\delta_{\text{Bc Thesis}} = -0.0022$, 95% CI $[-0.0146, 0.0103]$, $\delta_{\text{Internet}} = -0.0093$, 95% CI $[-0.0253, 0.0066]$, $\delta_{\text{Marathon}} = 0.0046$, 95% CI $[-0.0063, 0.0153]$, $\delta_{\text{Marathon-Manheim}} = -0.0036$, 95% CI $[-0.0177, 0.0105]$, and $\delta_{\text{Marathon-PhD}} = 0.0006$, 95% CI $[-00112., 0.0124]$.

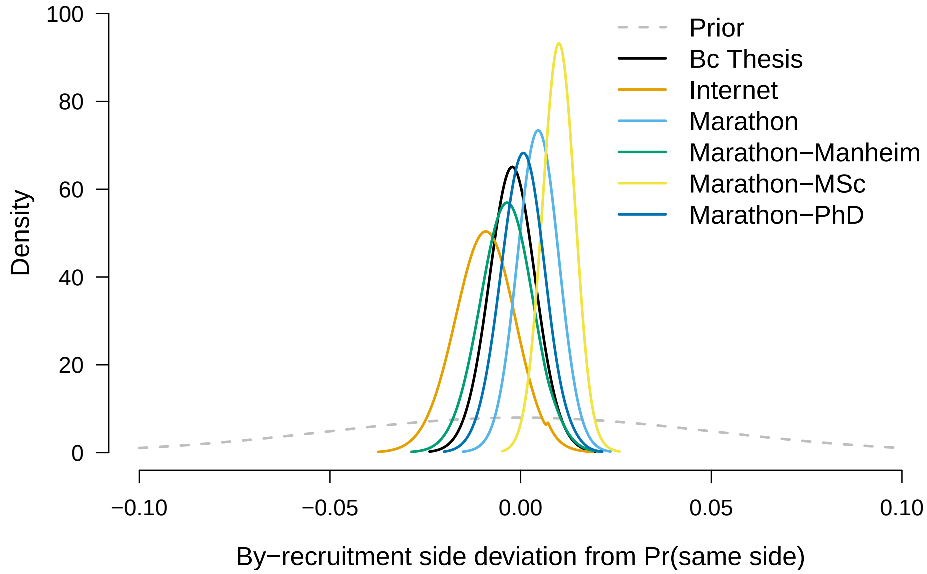

Figure 6: Posterior distribution of by-recruitment site differences from the probability of the same side show a considerable amount of uncertainty but no systematic differences. Prior distribution as grey dashed line, by-recruitment side posterior distributions in colors. Only the 95% CI for 'Marathon-MSc' does not contain zero; however, the difference disappears after excluding outliers, see Online Supplement Figure 3.

Removal of the four potential outliers (i.e., people with an observed probability of the same side larger than 0.53) resulted in a considerable decrease in the by-recruitment site differences from the probability of the same side. The 'Marathon-Msc''s 95% CI of by-recruitment site differences from the probability of the same side now contains zero, $\delta_{\text{Marathon-Msc}} = 0.0016$, 95% CI $[-0.0037, 0.0069]$, however, the by recruitment site difference for 'Marathon' no longer includes zero $\delta_{\text{Marathon}} = 0.0062$, 95% CI $[0.0004, 0.0120]$. See Figure 4 in Online Supplements for the corresponding visualization.

Summary

The analysis did not reveal reliable and systematic by-recruitment site differences from the same-side bias; 'Marathon-Msc' site deviation from the same-side bias's 95% CI did not contain zero using the complete dataset, and 'Marathon' site deviation from the same-side bias's 95% CI did not contain zero using the potential outliers removed data set. The posterior estimates were however considerably variable and a much larger data set would be needed to draw strong conclusions.

Appendix C: Frequentist results

At the request of editors and reviewers, we reproduce the main findings from our manuscript using a frequentist methodology. All CIs reported in this section correspond to 95% confidence intervals. In each subsection, we report results from an analysis performed on the full data set and follow-up results performed on the data set after excluding potential outliers (i.e., the four people with an observed probability of the same side larger than 0.53). All mixed-effects models were estimated using the `lme4` `R` package [version 1.1.35.3, 37].

Analyses assuming independent observations

Assuming independence of the individual coin tosses, a simple (nonhierarchical) binomial test finds a statistically significant effect of the same-side bias (178,079 same-side landings from 350,757 tosses), $p < 0.001$, Pr(same side) $=$

0.5077, 95% CI [0.5060, 0.5094]. Furthermore, a simple binomial test does not find a statistically significant effect of a heads-tails bias (175,421 heads out of 350,757 tosses), $p = 0.887$, Pr(heads) $= 0.5001$, 95% CI [0.4985, 0.5018].

Removing potential outliers results in a same-side bias that is smaller, Pr(same side) $= 0.5060$, 95% CI [0.5043, 0.5077] (171,517 same-side landings from 338,985 tosses), but remains statistically significant, $p < 0.001$. The probability of heads remains practically identical, Pr(heads) $= 0.5004$, 95% CI [0.4987, 0.5021] (169,635 heads out of 338,985 tosses), and statistically non-significant, $p = 0.626$.

Analyses accounting for dependencies due to people and coins

To account for the possible dependency between coin people, we specified a generalized linear mixed-effect model with a binomial likelihood and a logit link for the probability of a coin landing on the same side it started with: the intercept (b_μ, corresponding to the same-side bias), by-participant random intercepts (τ_{b_μ}, corresponding to the by-participant variability in the same-side bias), and the fixed-effect of starting side (b_1, a factor with the following levels: 'heads', 'tails'). The model reveals a statistically significant same-side bias $b_\mu = 0.0399$, se $= 0.0099$, $z = 4.03$, $p < 0.001$ (corresponding to the probability of the same side Pr(same side) $= 0.5100$, 95% CI [0.5051, 0.5148]) and a statistically non-significant effect of the starting side, $b_1 = 0.0007$, se $= 0.0034$, $z = 0.22$, $p = 0.829$. The model also reveals a statistically significant between-people variability in the same-side bias, $\tau_{b_\mu} = 0.0626$, $\chi^2(1) = 120.43$, $p < 0.001$ (by a comparison to a model without the random intercept). We could not test for a between-coin variance in the heads-tails bias due to the limited flexibility of lme4. However, we tested for a potential between-coin variance in the same-side bias—the model extension with the by-coin intercept in the same-side bias turns out statistically non-significant, $\chi^2(1) = 1.19$, $p = 0.276$.

Removing potential outliers does not qualitatively affect the results. The degree of the same-side bias decreases but remains statistically significant $b_\mu = 0.0241$, se $= 0.0057$, $z = 4.24$, $p < 0.001$ (corresponding to the probability of the same side Pr(same side) $= 0.5060$, 95% CI [0.5032, 0.5088]) and the effect of the starting side remains statistically non-significant, $b_1 = 0.0018$, se $= 0.0034$, $z = 0.52$, $p = 0.604$. The model again reveals a smaller but statistically significant between-people variability in the same-side bias, $\tau_{b_\mu} = 0.0278$, $\chi^2(1) = 36.05$, $p < 0.001$ (by a comparison to a model without the random intercept), and statistically non-significant extension with a between-coin variance in the same-side bias, $\chi^2(1) = 2.77$, $p = 0.096$.

We further refer the reader to the `dat.bartos2023` entry in the `metadat` [version 1.3-0, 38] `R` package by Wolfgang Viechtbauer, who re-analyzed the data set with a frequentist meta-analyses using the `metafor` [39] `R` package and arrived at qualitatively identical conclusions.

Learning effects

To estimate the learning effects, we extended the previously specified generalized linear mixed-effect model by including the linear (β_t) and quadratic (β_{t^2} fixed-effect of the number of coin flips (scaled by 10,000 to improve convergence), and the corresponding by-participant random slopes (τ_{b_t} and $\tau_{b_{t^2}}$; and the corresponding correlation between random-effects). This model allows us to estimate and test the initial same-side bias (the intercept b_μ) but the polynomial form does not provide an estimate of the asymptotic same-side bias. We first tested linear fixed- and random-effects of time and found that the model fit significantly improves upon the previously specified linear mixed-effect model, $\chi^2(3) = 45.42$, $p < 0.001$. Then we tested additional quadratic fixed- and random-effects of time and found the model fit significantly improves upon the linear fixed- and random-effects only model, $\chi^2(4) = 23.47$, $p = 0.001$. We did not pursue more complex models and report the linear and quadratic fixed and random-effects model. The model reveals a statistically significant initial same-side bias much larger than the previous models, $b_\mu = 0.0677$, se $= 0.0171$, $z = 3.96$, $p < 0.001$ and substantial between person variability in the initial same-side bias $\tau_{b_\mu} = 0.1048$. The statistically significant linear fixed-effect of time, $b_t = -0.1114$, se $= 0.0454$, $z = -2.45$, $p = 0.014$, and statistically non-significant quadratic fixed-effect of time, $b_{t^2} = 0.0505$, se $= 0.0305$, $z = 1.66$, $p = 0.098$, are both accompanied by notably between person variability in the corresponding effects, $\tau_{b_t} = 0.2196$ and $\tau_{b_{t^2}} = 0.1241$. The time effects result in a high initially probability of the same side, Pr(initial same side) $= 0.5169$, 95% CI [0.5085, 0.5253], which decreases quickly in the initial couple thousands flips and eventually starts increasing (with an extreme uncertainty) due to the quadratic form of the time effect (Figure 7). The results again confirm the statistically non-significant effect of the starting side, $b_1 = 0.0008$, se $= 0.0034$, $z = 0.23$, $p = 0.816$.

Removing the four potential outliers does not qualitatively affect the results. The linear fixed- and random-effects of time model significantly improves upon the simple linear mixed-effect model, $\chi^2(3) = 23.39$, $p < 0.001$; the quadratic fixed- and the random-effects of time model significantly improves upon the linear fixed- and random-effects of time model, $\chi^2(4) = 12.54$, $p = 0.014$. The statistically significant initial same-side bias decreases after excluding the potential outliers, $b_\mu = 0.0443$, se $= 0.0131$, $z = 3.37$, $p < 0.001$ as well as the between person variability in the

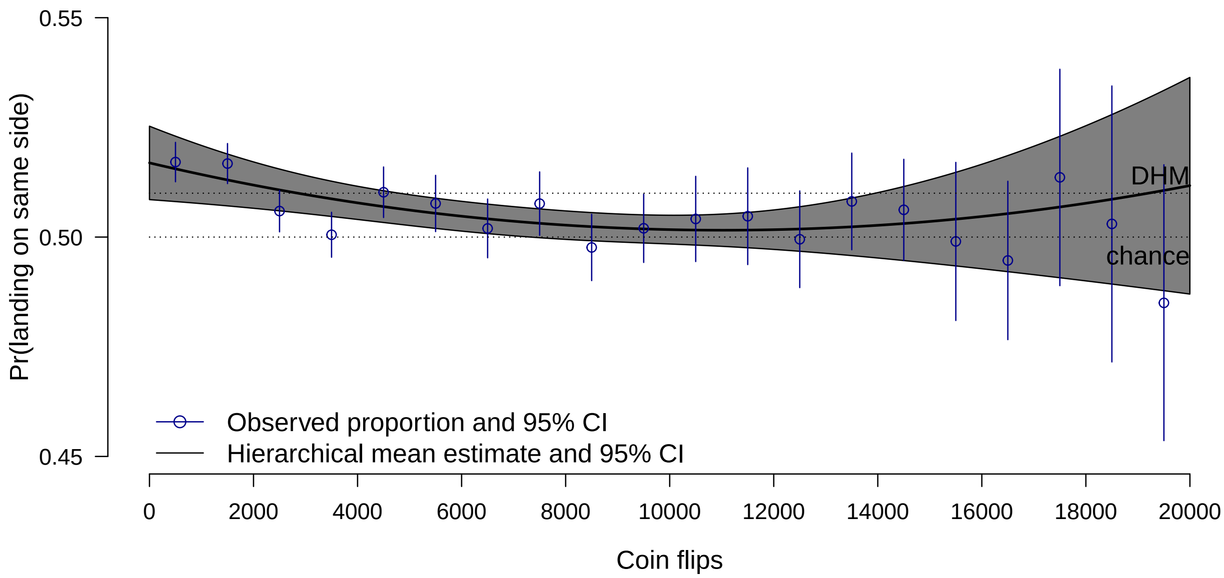

Figure 7: Frequentist analysis verifying the degree of the individual same-side bias decreases as people flip more coins. Mean estimate and 95% CI of the probability of the same side with the increasing number of coin flips from a generalized linear mixed effect model with binomial likelihood, logit link function, and linear and quadratic effect of time on the probability of the same side, vs. the observed proportion and 95% CI of the same side probability aggregated per sequences of $1{,}000$ coin flips.

initial same-side bias $\tau_{b_\mu} = 0.0683$. The smaller statistically non-significant linear fixed-effect of time, $b_t = -0.0613$, se $= 0.0387$, $z = -1.59$, $p = 0.113$, and the smaller statistically non-significant quadratic fixed-effect of time, $b_{t^2} = 0.0233$, se $= 0.0271$, $z = 0.86$, $p = 0.390$, are still accompanied by smaller but notable between person variability in the corresponding effects, $\tau_{b_t} = 0.1470$ and $\tau_{b_{t^2}} = 0.0877$. However, the time effects still result in smaller yet high initially probability of the same side, Pr(initial same side) $= 0.5111$, 95% CI $[0.5046, 0.5175]$, which still decreases quickly in the initial couple thousands flips and eventually starts increasing (with an extreme uncertainty) due to the quadratic form of the time effect (Figure 5 in Online Supplements). The results again confirm the statistically non-significant effect of the starting side, $b_1 = 0.0019$, se $= 0.0034$, $z = 0.54$, $p = 0.588$.

Same-side bias variability by recruitment site

Finally, we tested for the potential variability in the same-side bias according to the source of participant's recruitment. We specified a generalized linear mixed-effect model with a binomial likelihood and a logit link for the probability of a coin landing on the same side it started with: the intercept (b_μ, corresponding to the same-side bias), by-participant random intercepts (τ_{b_μ}, corresponding to the by-participant variability in the same-side bias), fixed-effects for the source of participants' recruitment ($b_{r1}, b_{r2}, \ldots, b_{r5}$; a factor with the following levels: 'Bc Thesis', 'Marathon', 'Internet', 'Marathon-Manheim', 'Marathon-MSc', 'Marathon-PhD'), and the fixed-effect of starting side (b_1, a factor with the following levels: 'heads', 'tails'). A comparison of the full model and a reduced model (omitting the the source of participants' recruitment) did not reveal a significant effect of the source of participants' recruitment, $\chi^2(5) = 6.65$, $p = 0.248$.

Removing potential outliers does not affect the results as the source of recruitment fixed effects remain statistically non-significant, $\chi^2(5) = 7.71$, $p = 0.173$. Figure 6 in Online Supplements visualizes the estimates.

Summary

All frequentist re-analyses are aligned with the Bayesian results presented in the manuscript; a statistically significant same-side bias, statistically non-significant heads-tails bias (with our without adjustment for between-people dependencies), statistically significant between-person heterogeneity in the same-side bias, and statistically significant learning effects that decrease the same-side bias over time. No results change after exclusion of potential outliers,

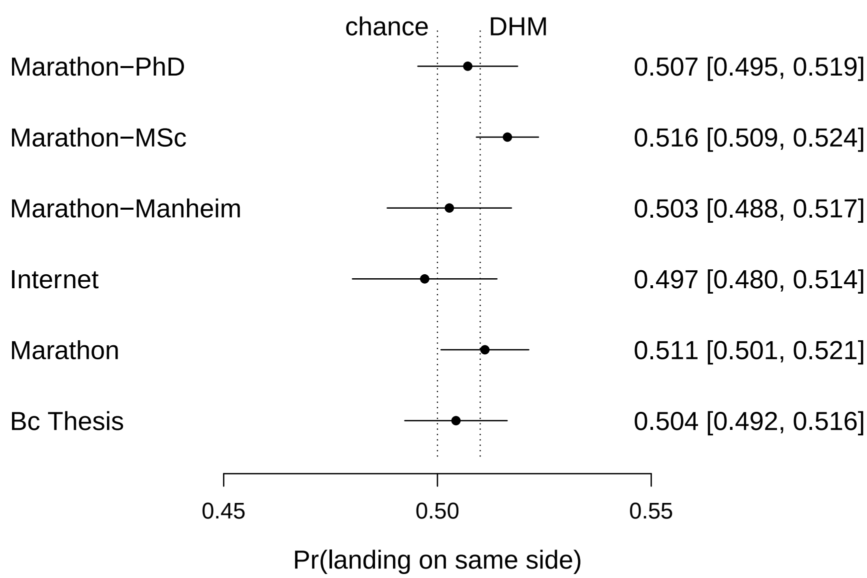

Figure 8: Frequentist marginal mean estimates and 95% CI of the probability of the same side do not differ by the source of participants' recruitment.