© 2018. Published by The Company of Biologists Ltd | Journal of Experimental Biology (2018) 221, jeb185454. doi:10.1242/jeb.185454

SHORT COMMUNICATION

Reanalysis of an oft-cited paper on honeybee magnetoreception reveals random behavior

Michael J. Baltzley[1,*] and Matthew W. Nabity[2]

ABSTRACT

While mounting evidence indicates that a phylogenetically diverse group of animals detect Earth-strength magnetic fields, a magnetoreceptor has not been identified in any animal. One possible reason that identifying a magnetoreceptor has proven challenging is that, like many research fields, magnetoreception research lacks extensive independent replication. Independent replication is important because a subset of studies undoubtedly contain false positive results and without replication it is difficult to determine whether the outcome of an experiment is a false positive. However, we report here a reanalysis of a well-cited paper on honeybee magnetoreception demonstrating that the original paper represented a false positive finding caused by incorrect estimates of probability. We also point out how good experimental design practices could have revealed the error prior to publication. Hopefully, this reanalysis will serve as a reminder of the importance of good experimental design in order to reduce the likelihood of publishing false positive results.

KEY WORDS: Magnetic field, *Apis mellifera*, Experimental design

INTRODUCTION

Despite evidence that magnetoreception is widespread in the animal kingdom, a magnetoreceptor and accompanying underlying neural circuitry have yet to be identified in any animal (Shaw et al., 2015; Clites and Pierce, 2017; Nordmann et al., 2017). Many reasons have been cited for why magnetoreceptors have proven elusive, including the absence of large, obvious magnetoreceptive organs and the possibility that magnetoreceptors could exist anywhere within an animal (Johnsen and Lohmann, 2005; Shaw et al., 2015). However, the field has also likely been hindered by the inevitability that a subset of the published findings on magnetoreception are false positives. In addition to published examples of failures to replicate specific magnetoreception studies (Klotz et al., 1997; Hert et al., 2011; Landler et al., 2018), there is increasing evidence that false positives are a wide-spread problem in published research (Ioannidis, 2005; Collins and Tabak, 2014; Open Science Collaboration, 2015). While replication can eventually lead to the identification of false positives, poor experimental design and misunderstanding of statistical analyses facilitate the publication of false positives, which can then cause other researchers to waste resources (Simmons et al., 2011). Recently, we discovered that the conclusions of Kirschvink et al. (1997), a well-cited article on honeybee (*Apis mellifera carnica*, Pollman 1879) magnetoreception, were based on incorrect data analyses.

According to the Journal of Experimental Biology website (http://jeb.biologists.org/content/200/9/1363.article-info; accessed 5 September 2018), Kirschvink et al. (1997) has been cited in the scientific literature over 40 times, or about twice per year since it was published. The article is still currently being cited as evidence for magnetoreception in bees (e.g. Prato et al., 2013; Ferrari, 2014; Pereira-Bomfim et al., 2015; Liang et al., 2016; Lambinet et al., 2017; Kong et al., 2018). However, the number of citations understates the impact of Kirschvink et al. (1997). The article abstract on the Journal of Experimental Biology website has been accessed over 2200 times since 2001, including 193 times in the first 8 months of 2018, while the full-text PDF has been accessed over 3200 times since 2001, including 158 times in the first 8 months of 2018.

Unfortunately, the positive findings of Kirschvink et al. (1997) rely solely on incorrect estimates of probability. The authors trained bees to use a magnetic field to distinguish between a positive reward (sucrose) and a negative reward (electric shock). Once the bees learned to associate the magnetic field cue with a positive reward, the authors reduced the magnetic field strength and allowed the bees to learn to associate the weaker stimulus with the food reward. If the bees succeeded in learning the new association, the magnetic field strength was reduced again. This process was continued until the bees could no longer learn to associate the magnetic field and the positive stimulus, presumably because they could no longer detect the magnetic field.

The error the authors made was in their criteria for determining whether or not the bees had learned to associate the magnetic field with the positive stimulus. The bees were given approximately 80 trials to make either six consecutive correct decisions or at least seven of eight correct decisions (i.e. seven of eight or eight of eight). The authors stated that there was only a 1.6% chance and a 3.5% chance, respectively, that bees would achieve these levels of success randomly. However, the authors failed to consider that over the course of 80 trials the bees had up to 75 opportunities to reach one of the learning criteria. The actual probability of reaching a learning criterion over the course of 80 trials if the bees were choosing targets randomly was approximately 66.5%. We were able to produce similar results to Kirschvink et al. (1997) using a random number generator, thereby demonstrating the fundamental flaw in the experimental design. Hopefully this example will encourage other researchers to consider their experimental design carefully before embarking on experiments.

MATERIALS AND METHODS

The probability of bees reaching a criterion (six correct choices in a row or seven out of eight choices) was determined using a random number generator in R (https://www.R-project.org/).

[1]Department of Biology, Western Oregon University, 345 Monmouth Avenue N., Monmouth, OR 97361, USA. [2]Department of Mathematics, Western Oregon University, 345 Monmouth Avenue N., Monmouth, OR 97361, USA.

*Author for correspondence (baltzlem@wou.edu)

M.J.B., 0000-0001-5789-4760

Received 28 May 2018; Accepted 24 September 2018

We used the function *rbinom* to create one-million 80-trial blocks of zeros and ones, then counted how many of the 80-trial blocks contained a sequence where the number one occurred in six of six trials or in seven of eight trials. We performed 10 replicates and found that randomly behaving 'bees' reached one of the established criteria (6 of 6 or 7 of 8) in 66.5±0.1% (mean±s.d.) of 80-trial blocks.

To confirm these results, we also determined the probability of bees reaching a criterion using numerical experiments carried out in Matlab R2017a (MathWorks, Natick, MA, USA). Eighty-trial blocks were modeled by arrays of 80 random numbers generated using the built-in functions *rand* and *randn*. The numbers generated by *rand* are uniformly distributed on [0, 1]. If the number was greater than 0.5, we set the value to one, otherwise we set the value to zero. Similarly, the numbers generated by *randn* are normally distributed on [−1, 1]; we set the value to one if it was greater than zero, otherwise we set the value to zero. The arrays were then analyzed to determine whether ones occurred in six of six trials or in seven of eight trials within each 80-trial block.

All random numbers in the Matlab simulations were generated using the Mersenne Twister random number generator. This random number generator is widely used and is also the default random number generator in R. We used the default seed and a controlled seed for comparison. Using *rand* and performing 500,000 80-trial blocks, at least one of the learning criteria was reached in 66.3% of the trials using the controlled seed and in 66.5% of the trials using the default seed (Table S1). Using *randn* and performing 500,000 80-trial blocks, a criterion was reached in 66.5% of trials using the controlled seed and in 66.4% of the trials using the default seed.

To reanalyze the data from Kirschvink et al. (1997), we determined the proportion of bees that reached one of the criteria at each magnetic field intensity by drawing a horizontal gridline through the appropriate data point in fig. 3 from Kirschvink et al. (1997) and identifying the *y*-intercept (Table 1; Fig. S1). Gridlines were drawn using Adobe Illustrator CS 5 (Adobe Systems Incorporated, San Jose, CA, USA). Kirschvink et al. (1997) performed two experiments: one experiment tested the ability of 15 bees to learn to associate a 10 Hz AC magnetic field with a sucrose reward, and the second experiment tested the ability of 11 bees to learn to associate a 60 Hz AC magnetic field with a sucrose reward. The proportions of successful bees as reported in Kirschvink et al. (1997) did not align exactly with the proportions that would be produced using 15 or 11 bees; therefore, our calculations for the number of successful bees were rounded to the nearest integer. For the experiment with 15 bees exposed to 10 Hz AC magnetic fields, the data point for 1300 μT fell almost exactly between two possible proportions, so the data were analyzed using both possible values for that particular data point.

Using an expected probability of 66.5%, we performed exact multinomial tests for both sets of data from Kirschvink et al. (1997) using the *XNomial* package in R (https://CRAN.R-project.org/package=XNomial). If 66.5% of bees randomly reach a learning criterion, 33.5% of bees would be expected to not reach a learning criterion for the first magnetic field stimulus. Of the 66.5% of bees that reached a criterion for the first magnetic field stimulus, 66.5% would be expected to reach a criterion for the second association, etc. It should be noted that in the original experiment by Kirschvink et al. (1997), bees were given approximately 80 attempts, and at least one trial was terminated early because a bee failed to return after 19 trials.

We also performed 15 simulations of individual bees using a random number generator in R. Mimicking the protocol in Kirschvink et al. (1997), we gave simulated bees up to 80 trials to make six of six or seven of eight correct choices; if a simulated bee reached one of the criteria, we concluded that the 'bee' had learned. The simulated bee then began a new set of trials and had up to an additional 80 trials to reach a criterion. This process continued until the simulated bee failed to reach one of the learning criteria.

Kirschvink et al. (1997) based their experimental design on that of Walker and Bitterman (1989). Walker and Bitterman (1989), however, used a DC magnetic field and, for most trials, allowed the bees 32 attempts to reach the learning criteria. We performed exact multinomial tests for the data from Walker and Bitterman (1989) using a predicted probability of success of 32.7%, which was determined using 10 random number simulations of one-million 32-trial blocks in R (32.7±0.04%; mean±s.d.).

In the Kirschvink et al. (1997) and Walker and Bitterman (1989) experiments, any bee that reached a criterion was tested again under a weaker magnetic field. Therefore, the data points for each magnetic field strength are not independent and represent pseudoreplication. For example, in their experiment with a 60 Hz AC magnetic field, Kirschvink et al. (1997) performed 25 trials with 11 bees; bees succeeded in reaching the learning criteria in 14 of those trials. To avoid the problem of non-independence, for the exact multinomial tests where we compared the original experimental data with the predicted random distribution, we only

Table 1. Data from Kirschvink et al. (1997) and predicted results if bees were randomly choosing targets

	Kirschvink et al. (1997)				Predicted results		
	10 Hz field 15 bees		60 Hz field 11 bees			Simulation for 15 'bees'	
Magnetic field strength	Proportion	No. of bees	Proportion	No. of bees	66.5% chance of success	No. of bees	Proportion
No learning	0.30	5 or 4*	0.35	4	0.34	4	0.27
1300 μT	0.70	10 or 11*	0.65	7	0.67	11	0.73
430 μT	0.59	9	0.34	4	0.44	8	0.53
130 μT	0.38	6	0.26	3	0.29	6	0.40
43 μT	0.31	5	0.00	0	0.20	5	0.33
13 μT	0.25	4			0.13	5	0.33
4.3 μT	0.22	3			0.09	5	0.33
1.3 μT	0.08	1			0.06	3	0.20
430 nT	0.07	1			0.04	1	0.07
130 nT	0.00	0			0.03	1	0.07

*For their experiment with 10 Hz AC magnetic fields, Kirschvink et al. (1997) used 15 bees. Statistical tests were performed using both 10 bees and 11 bees for 1300 μT magnetic fields because the provided data indicated that 10.5 bees (0.70×15) were able to learn to associate the reward with a 1300 μT magnetic field stimulus.

Journal of Experimental Biology (2018) 221, jeb185454. doi:10.1242/jeb.185454

used the lowest field strength each bee was reported to detect. For example, using 11 bees and a 60 Hz AC magnetic field, Kirschvink et al. (1997) reported that 4 bees failed to reach a criterion, 7 succeeded when tested with a 1300 µT magnetic field, 4 succeeded at 430 µT, 3 succeeded at 130 µT and 0 succeeded at 43 µT (Table 1). For our analysis, we used the following values: 4 bees failed to reach a criterion, 3 bees did not succeed below 1300 µT, 1 bee did not succeed below 430 µT, 3 bees did not succeed below 130 µT and 0 bees succeeded at 43 µT (Table S2).

For our experiment with 15 simulated bees, some of the magnetic field categories had values of zero (e.g. all 5 simulated bees that reached a criterion at 43 µT also reached a criterion at 13 µT). Because exact multinomial tests cannot be performed with expected values of zero, we chose to use all available data points for our statistical analysis, even though the data points were not independent.

The recreated data from Kirschvink et al. (1997) and Walker and Bitterman (1989) with the results of the statistical analyses, as well as simulated data sets are available at https://figshare.com/s/87a60291069dddb35910.

RESULTS AND DISCUSSION

Kirschvink et al. (1997) stated that the probability of bees reaching the learning criterion of six correct choices in a row was 1.6%, and the probability of making seven or eight correct choices out of eight attempts was 3.5%. However, because the bees were given multiple attempts to reach the criteria, the probability of reaching the criteria was much higher (Fig. 1A). In fact, the probability of a bee reaching the given criteria randomly was greater than 5% after only nine trials. We found that the data from the Kirschvink et al. (1997) experiment with bees exposed to 10 Hz AC magnetic fields were not significantly different from the expected outcome if the bees were choosing targets randomly, regardless of whether 10 or 11 bees succeeded at the first magnetic field strength (exact multinomial test: 10 bees, P=0.17; 11 bees, P=0.23; Fig. 1B, Table 1). Likewise, the data from the Kirschvink et al. (1997) experiment with bees exposed to 60 Hz AC magnetic fields were not significantly different from the expected outcome if the bees were choosing targets randomly (exact multinomial test: P=0.17).

The results of our random number simulation with 15 'bees' were also not significantly different from the data of Kirschvink et al. (1997) (exact multinomial test: 10 bees, P=0.96; 11 bees, P=0.98). An example 'bee' from our simulation is shown alongside a recreation of fig. 2 from Kirschvink et al. (1997) for comparison (Fig. 2A,B). For simulated bees that reached a criterion, the average number of trials it took to reach the criterion was 28.6±15.4 (mean± s.d.). Some of our simulated bees showed a pattern of results that looked like they were learning, while others showed a pattern that looked like they were making random choices (Fig. 2C).

The reasonable conclusion to make from the results published in Kirschvink et al. (1997) is that the bees did not learn to associate either 10 Hz AC or 60 Hz AC magnetic fields with a positive reward. Kirschvink et al. (1997) should no longer be cited as evidence that bees can detect magnetic fields.

In addition to the incorrect estimate of probability, there were several other experimental design concerns in Kirschvink et al. (1997). First, because bees were removed from testing as soon as they failed to reach a criterion, each subsequent field strength was presented to a smaller number of bees. As a result, the proportion of bees that succeeded inevitably continued to decrease, thereby creating the appearance of a dose–response curve even if the success of a given bee occurred simply by random chance. All possible experimental outcomes, other than 0% success or 100% success, would have created the appearance that the response of the bees decreased as the magnetic field intensity decreased.

A second problem with the experimental design was that although Kirschvink et al. (1997) did not know which target was the reward until after a given trial, they were aware of the outcome of a trial prior to the initiation of the next trial and prior to data analysis;

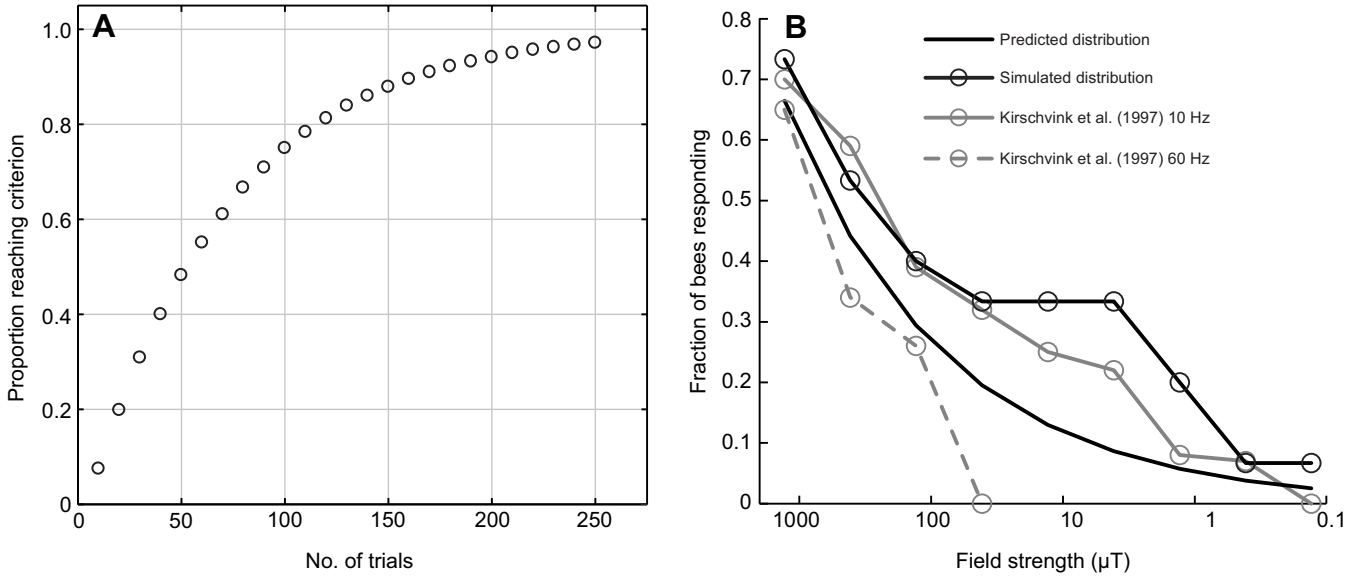

Fig. 1. Actual proportion of bees reaching the learning criteria (6 of 6 or 7 of 8 correct choices) and predicted proportion based on random number generation. (A) The proportion of bees expected to reach a criterion increases as the number of opportunities to reach the criteria increases. For each data point, the proportion reaching a criterion was determined using 250,000 blocks created using the *rand* function and default seed in Matlab (see Materials and Methods). (B) Predicted and actual proportions of honeybees able to discriminate each magnetic field stimulus tested in Kirschvink et al. (1997). The data from Kirschvink et al. (1997) for 10 Hz AC magnetic fields represent 15 bees, while those for 60 Hz AC magnetic fields represent 11 bees. The predicted distribution is based on the probability that 66.5% of bees, if they are choosing targets randomly, will reach the criteria for learning for each magnetic field stimulus level. The simulated distribution (15 bees) was created using a random number generator. The distributions from Kirschvink et al. (1997) were not significantly different from either the predicted distribution or the simulated distribution.

Journal of Experimental Biology (2018) 221, jeb185454. doi:10.1242/jeb.185454

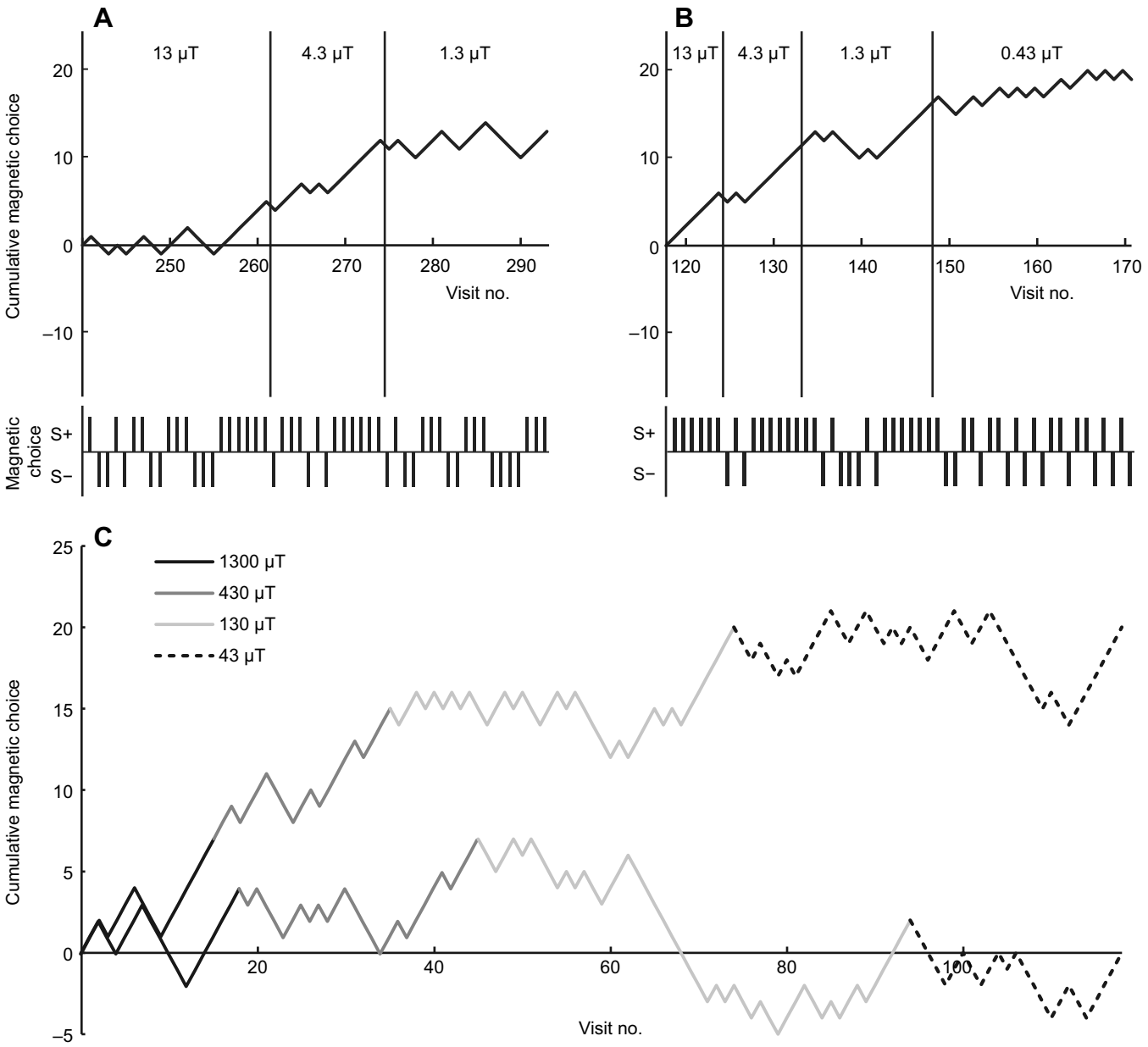

Fig. 2. Example data of real and simulated honeybees learning to associate a 10 Hz AC magnetic field with a sucrose reward. (A) Data from fig. 2 of Kirschvink et al. (1997). The bottom panel shows the choice of the bee at each visit, where S+ indicates that the target associated with the magnetic field, and ultimately a sucrose reward, was chosen. The top panel shows the cumulative S+ and S– choices, with vertical lines indicating that the bee reached a learning criterion (6 of 6 or 7 of 8 correct choices), at which time the magnetic field was reduced. (B) Data for a simulated bee created with a random number generator. Vertical lines indicate that the 'bee' reached a criterion. (C) Example data of two simulated bees learning to associate a 10 Hz AC magnetic field with a sucrose reward. Changes in the color of the line indicate when a 'bee' reached a criterion. Using a random number generator, strings of positive choices that meet a criterion occur frequently. By intentional selection of which example to display, it is possible to make a 'bee' look like it is learning (top line) or to make a 'bee' look like it is performing a random series of choices.

therefore, the experiments were not performed blind. As a result of this design, the researchers stopped any given 80-trial block early if the expected outcome was observed, but allowed the experiment to continue if the expected outcome was not observed. The researchers should have continued the experiment through 80 trials and then examined whether or not the bees continued to perform above random chance once they had reached the learning criteria.

An additional problem with the experimental design was that there were no experimental controls to determine how bees behaved in the absence of a magnetic field. The use of experimental controls would have likely revealed the incorrect estimation of probability. Finally, because the magnetic field stimuli were not presented in a random order, the magnetic field effects and temporal effects were confounded. Proper randomization would have shown that the bees were randomly choosing targets rather than learning a discrimination task.

Kirschvink et al. (1997) was an important paper because it was a rare example of the independent replication of a previous experiment that demonstrated magnetoreception in bees (Walker and Bitterman, 1989; Vácha and Soukopová, 2004). There were two primary differences between the experimental designs described in Walker and Bitterman (1989) and Kirschvink et al. (1997). Walker

and Bitterman (1989) only allowed the bees approximately 32 trials to reach one of the learning criteria. If bees were choosing randomly, a criterion would be expected to be reached in 33% of trials. The other difference was that Kirschvink et al. (1997) used an automated delivery system so that the reward was not made available until after the bees made a choice, whereas in Walker and Bitterman (1989), the bees could have potentially smelled the difference between the positive and negative reward. While Walker and Bitterman (1989) made the same incorrect estimates of probability, and had similar experimental design flaws, our reanalysis of the Walker and Bitterman (1989) data found that their results were significantly different from the predicted distribution if bees were choosing targets randomly (Table S2; exact multinomial test: $P<0.000001$). Based on the median performance of their bees, Walker and Bitterman (1989) stated that the threshold of magnetic field intensity detection was 260 nT; however, because of the above concerns regarding experimental design, this conclusion should be reconsidered. To our knowledge, this particular protocol has not been used in other studies of magnetoreception in bees.

Experimental design problems are not uncommon in biological research (Holman et al., 2015). Concerns about the reproducibility of results have also gained significant attention in recent years, particularly in medical research and in psychology research (Begley and Ellis, 2012; Open Science Collaboration, 2015; Baker, 2016; Johnson et al., 2017). False positives cannot be eliminated, but they can be reduced by proper experimental design including randomization, *a priori* determination of statistical analyses to be performed, blind data collection and independent evidence of an error before an outlier is discarded (Festing and Altman, 2002; van Wilgenburg and Elgar, 2013; Holman et al., 2015; Curtis et al., 2015). While the example we presented here is from the field of magnetoreception research, we hope it will serve as a valuable reminder for all experimental biologists to both carefully consider their own experimental design and critically evaluate the methods within published research studies.

Acknowledgements
We thank P. Aldrich and two anonymous reviewers for providing valuable comments on the manuscript.

Competing interests
The authors declare no competing or financial interests.

Author contributions
Conceptualization: M.J.B., M.W.N.; Methodology: M.J.B., M.W.N.; Software: M.J.B., M.W.N.; Validation: M.W.N.; Formal analysis: M.J.B.; Data curation: M.J.B.; Writing - original draft: M.J.B.; Writing - review & editing: M.J.B., M.W.N.; Visualization: M.J.B., M.W.N.

Funding
This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Data availability
Data are available from the figshare repository: https://figshare.com/s/87a60291069dddb35910

Supplementary information
Supplementary information available online at http://jeb.biologists.org/lookup/doi/10.1242/jeb.185454.supplemental

References
Baker, M. (2016). 1,500 scientists lift the lid on reproducibility *Nature* **533**, 452-454.
Begley, C. G. and Ellis, L. M. (2012). Drug development: raise standards for preclinical cancer research. *Nature* **483**, 531-533.
Clites, B. L. and Pierce, J. T. (2017). Identifying cellular and molecular mechanisms for magnetosensation. *Annu. Rev. Neurosci.* **40**, 231-250.
Collins, F. S. and Tabak, L. A. (2014). NIH plans to enhance reproducibility. *Nature* **505**, 612-613.
Curtis, M. J., Bond, R. A., Spina, D., Ahluwalia, A., Alexander, S. P. A., Giembycz, M. A., Gilchrist, A., Hoyer, D., Insel, P. A., Izzo, A. A. et al. (2015). Experimental design and analysis and their reporting: new guidance for publication in BJP. *Br. J. Pharmacol.* **172**, 3461-3471.
Ferrari, T. E. (2014). Magnets, magnetic field fluctuations and geomagnetic disturbances impair the homing ability of honey bees (*Apis mellifera*). *J. Apicult. Res.* **53**, 452-465.
Festing, M. F. W. and Altman, D. G. (2002). Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J.* **43**, 244-258.
Hert, J., Jelinek, L., Pekarek, L. and Pavlicek, A. (2011). No alignment of cattle along geomagnetic field lines found. *J. Comp. Physiol. A.* **197**, 677-682.
Holman, L., Head, M. L., Lanfear, R. and Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biol.* **13**, e1002190.
Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med.* **2**, e124.
Johnsen, S. and Lohmann, K. J. (2005). The physics and neurobiology of magnetoreception. *Nat. Rev. Neurosci.* **6**, 703-712.
Johnson, V. E., Payne, R. D., Wang, T., Asher, A. and Mandal, S. (2017). On the reproducibility of psychological science. *J. Am. Stat. Assoc.* **112**, 1-10.
Kirschvink, J., Padmanabha, S., Boyce, C. and Oglesby, J. (1997). Measurement of the threshold sensitivity of honeybees to weak, extremely low-frequency magnetic fields. *J. Exp. Biol.* **200**, 1363-1368.
Klotz, J. H., Van Zandt, L. L., Reid, B. L. and Bennett, G. W. (1997). Evidence lacking for magnetic compass orientation in fire ants (Hymenoptera: Formicidae). *J. Kansas Entomol. Soc.* **70**, 64-65.
Kong, L.-J., Crepaz, H., Górecka, A., Urbanek, A., Dumke, R. and Paterek, T. (2018). *In-vivo* biomagnetic characterisation of the American cockroach. *Sci. Rep.* **8**, 5140.
Lambinet, V., Hayden, M. E., Reigl, K., Gomis, S. and Gries, G. (2017). Linking magnetite in the abdomen of honey bees to a magnetoreceptive function. *Proc. Biol. Sci.* **284**, 20162873.
Landler, L., Nimpf, S., Hochstoeger, T., Nordmann, G. C., Papadaki-Anastasopoulou, A. and Keays, D. A. (2018). Comment on "Magnetosensitive neurons mediate geomagnetic orientation in *Caenorhabditis elegans*". *eLife* **7**, e30187.
Liang, C.-H., Chuang, C.-L., Jiang, J.-A. and Yang, E.-C. (2016). Magnetic sensing through the abdomen of the honey bee. *Sci. Rep.* **6**, 23657.
Nordmann, G. C., Hochstoeger, T. and Keays, D. A. (2017). Magnetoreception—a sense without a receptor. *PLoS Biol.* **15**, e2003234.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science* **349**, aac4716.
Pereira-Bomfim, M. G. C., Antonialli-Junior, W. F. and Acosta-Avalos, D. (2015). Effect of magnetic field on the foraging rhythm and behavior of the swarm-founding paper wasp *Polybia paulista* Ihering (Hymenoptera: Vespidae). *Sociobiology* **62**, 99-104.
Prato, F. S., Desjardins-Holmes, D., Keenliside, L. D., DeMoor, J. M., Robertson, J. A. and Thomas, A. W. (2013). Magnetoreception in laboratory mice: sensitivity to extremely low-frequency fields exceeds 33 nT at 30 Hz. *J. R. Soc. Interface* **10**, 20121046.
Shaw, J., Boyd, A., House, M., Woodward, R., Mathes, F., Cowin, G., Saunders, M. and Baer, B. (2015). Magnetic particle-mediated magnetoreception. *J. R. Soc. Interface* **12**, 20150499.
Simmons, J. P., Nelson, L. D. and Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22**, 1359-1366.
Vácha, M. and Soukopová, H. (2004). Magnetic orientation in the mealworm beetle Tenebrio and the effect of light. *J. Exp. Biol.* **207**, 1241-1248.
van Wilgenburg, E. and Elgar, M. A. (2013). Confirmation bias in studies of nestmate recognition: a cautionary note for research into the behaviour of animals. *PLoS ONE* **8**, e53548.
Walker, M. M. and Bitterman, M. E. (1989). Honeybees can be trained to respond to very small changes in geomagnetic field intensity. *J. Exp. Biol.* **145**, 489-494.

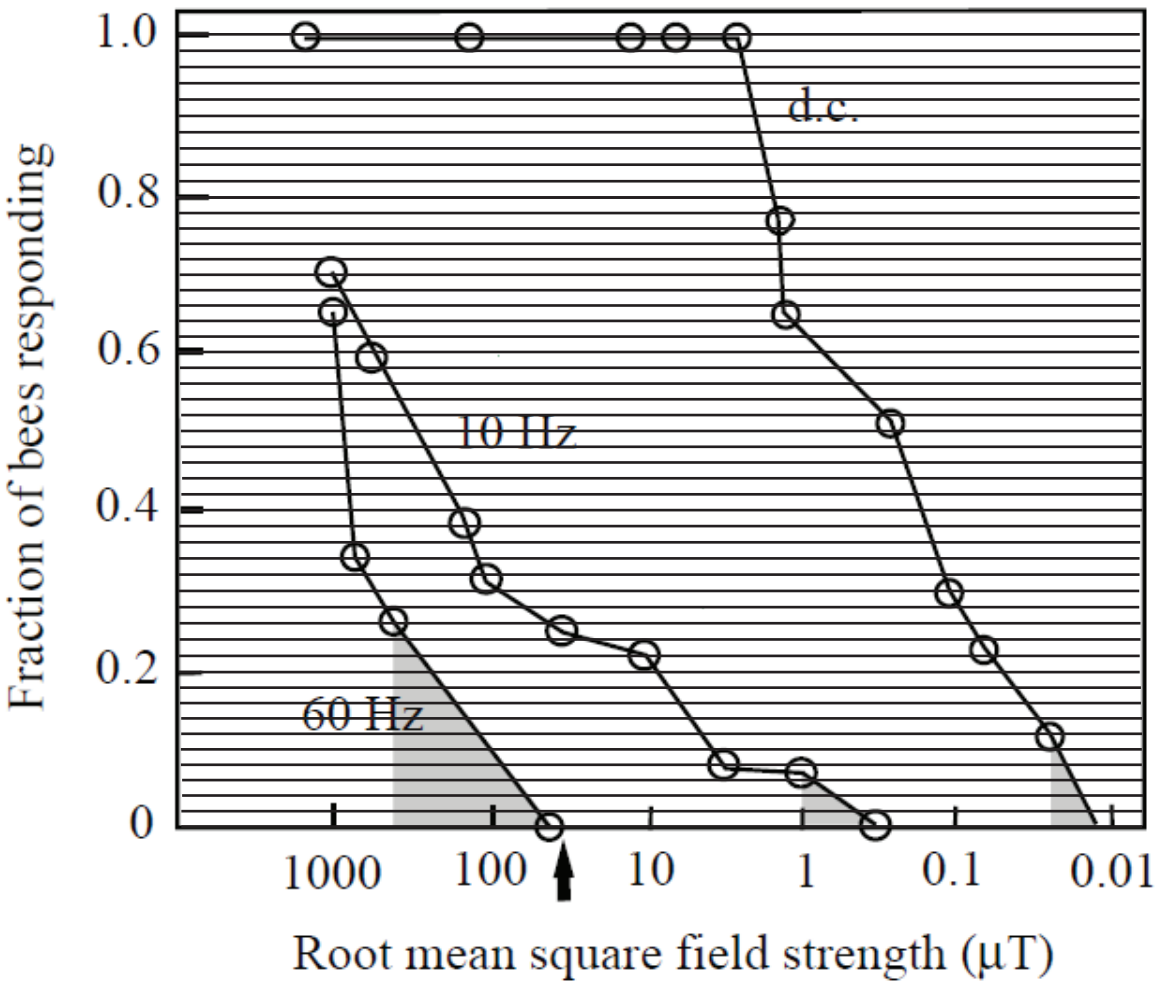

Fig. S1: Fig. 3 from Kirschvink *et al*. (1997) with gridlines used to recreate the raw data.

Table S1: Outcomes of experiments in Matlab with 80-trial blocks of ones and zeros created with two different random number generators and two different seeds. Both learning criteria could be met by a sequence of numbers such as 10111111.

Function	Uniform	Number of 80-trails blocks with:				Blocks with criteria met
		6 of 6	7 of 8	Both 6 of 6 and 7 of 8	Criteria not met	
rand	Controlled Seed	105,995	176,963	48,785	168,257	66.35%
	Default Seed	106,050	177,722	48,738	167,490	66.50%
randn	Controlled Seed	106,190	177,687	48,589	167,534	66.49%
	Default Seed	106,091	177,332	48,631	167,946	66.41%

Table S2: Summary of data used for exact multinomial tests.

	Magnetic field strength (μT)										
	None	1300	430	130	100	43	13	4.3	1.3	0.43	Total
Kirschvink *et al.* (1997), 10 Hz ac field, 15 bees; expected and simulated results were generated using a 66.5% probability of success											
Proportion from Kirsch. *et al.*, Fig.3	0.30	0.70	0.59	0.38	0.31	0.25	0.22	0.08	0.07	0.00	
Equivalent number of bees that reached criterion	4 or 5	11 or 10	9	6	5	4	3	1	1	0	44
Lowest field strength where criterion was reached	4 or 5	2 or 1	3	1	1	1	2	0	1	0	15
Expected proportion for lowest field strength reached	0.34	0.22	0.15	0.10	0.07	0.04	0.03	0.02	0.01	0.02	
15 'bee' simulation: 'bees' that reached criterion at each field strength	4	11	8	6	5	5	5	3	1	1	49
15 'bee' simulation: lowest field strength where criterion was reached	4	3	2	1	0	0	2	2	0	1	15
Kirschvink *et al.* (1997), 60 Hz ac field, 11 bees; expected results were generated using a 66.5% probability of success											
Proportion from Kirsch. *et al.*, Fig. 3	0.35	0.65	0.34	0.26	0.00						
Equivalent number of bees that reached criterion at each field strength	4	7	4	3	0						18
Lowest field strength where criterion was reached	4	3	1	3	0						11
Expected proportion for lowest field strength reached	0.34	0.22	0.15	0.10	0.20						

	Magnetic field strength (μT)												
	None	1200	120	12	5.6	2.6	1.2	0.56	0.26	0.12	0.056	0.026	0.012	Total
Walker and Bitterman (1989), dc field, 9 bees; expected results were generated using a 32.7% probability of success														
Number of bees that reached criterion at each field strength	0	9	9*	9	9	9	7	6	5	3	2	1	0	70
Lowest field strength where criterion was reached	0	0	0	0	0	2	1	1	2	1	1	1	0	9
Expected proportion for lowest field strength reached	0.67	0.22	0.07	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	

*For 5 of the 9 bees, Walker and Bitterman also tested the bees using magnetic field strengths of 56 μT and 26 μT; we did not include these data points in our analysis.