Human Ethology Bulletin 32 (2017)1- Proc. of the XXIII Biennial ISHE Congress: 53-59
Brief Report

BASELINE PROBABILITIES FOR TWO-ALTERNATIVE FORCED CHOICE TASKS WHEN JUDGING STIMULI IN EVOLUTIONARY PSYCHOLOGY: A METHODOLOGICAL NOTE

Thomas Pollet [1,2] & Anthony Little [3]

[1] Experimental and Applied Psychology, Vrije Universiteit Amsterdam, Amsterdam, Netherlands
[2] Social and Organisational Psychology, Universiteit Leiden, Leiden, Netherlands
[3] Department of Psychology, University of Bath, Bath, United Kingdom

t.v.pollet1981@gmail.com

ABSTRACT

Many paradigms in evolutionary psychology involve forced choice tasks with two alternatives. While the number of trials used across studies varies substantially, in such tasks it is common to test against a baseline of 50% (often via a one-sample t-test). In this paper, we simulate forced choice designs, varying in sample sizes (30 to 120) and number of trials (2 to 34) to empirically examine the usefulness of a 50% benchmark. Our results show that 50% is a weak benchmark when using a small number of trials. The simulations also indicate that increasing the number of trials is beneficial if one wants to use a 50% benchmark. There are however, marginal returns to increasing the number of trials: moving from 2 to 8 trials matters substantially more than moving from 28 to 34. Our approach also illustrates the value of simulations for understanding experimental designs, such as forced choice tasks, in evolutionary psychology.

Keywords: *Simulations, methodology, probability, base rate.*

INTRODUCTION

Many paradigms in evolutionary psychology, and in fact many other disciplines, employ a forced choice design whereby participants choose from two alternatives in a stimulus

https://doi.org/heb/10.22330/321/053-059

set, often termed two-alternative forced choice (2AFC) tasks. Examples of topics studied via these paradigms include electoral decision making (e.g., Little, Burriss, Jones, & Roberts, 2007), facial attractiveness (e.g., DeBruine, 2004; Perrett et al., 1998; Rhodes, Proffitt, Grady, & Sumich, 1998; Roberts et al., 2004), masculinity/femininity (e.g., Penton-Voak & Chen, 2004), selecting characteristics of prospective mates (e.g., Bressler & Balshine, 2006; Haselton & Miller, 2006; Li & Kenrick, 2006), jealousy (e.g., Bendixen, Kennair, & Buss, 2015; Buller, 2005; Buss, Larsen, Westen, & Semmelroth, 1992; Schützwohl, 2004), cheating and trust (e.g., Stirrat & Perrett, 2010; Verplaetse, Vanneste, & Braeckman, 2007), moral decision making (e.g., Bleske-Rechek, Nelson, Baker, Remiker, & Brandt, 2010; Kurzban, DeScioli, & Fein, 2012), and intentions to act (e.g., Barrett, Todd, Miller, & Blythe, 2005). There is a wide variation in how many forced choice trials are used from just a single trial (e.g., Buller, 2005 on jealousy dilemmas) to several dozen (e.g., Penton-Voak & Chen, 2004).

In terms of analyses, the data from these 2AFC experiments are often tested against a pre-determined 50% chance level (after pooling). For example, researchers will have participants complete eight forced choice trials, average the response across trials for each participant (e.g., 4 out of 8 or 0.5 or 50%) and then test that proportion against chance (50%) (e.g., Little et al., 2007). In some instances, however, researchers have also calculated measures derived from signal detection theory (Macmillan, 2002). This paper does not deal with those cases and instead exclusively focuses on the use of a 50% benchmark, as it remains common practice to use such a cut-off and test against this. Here we simulate these types of designs and examine how well 50% represents an adequate benchmark at varying trial numbers and sample sizes.

It must also be noted that in many cases, researchers do not just rely on forced choices but use Likert ratings as well (e.g., Bressler & Balshine, 2006; Tovée, Edmonds, & Vuong, 2012). We do not cover the choice of one design versus another but exclusively focus on the use of a 50% benchmark in a forced choice paradigm. Also, this paper is not intended as an introduction into statistical simulations (e.g., Stulp & Barrett, 2013). Finally, the purpose of this paper is explicitly not to discuss the potential problems of pooling data across trials, which have been discussed extensively elsewhere (e.g., Kievit et al., 2013; Pollet et al., 2015). Instead, the purpose of this short methodological note is to derive the 'true' probability for experiments varying in sample size and number of trials. Moreover, this will inform researchers as to how many participants and trials they should obtain/use if they want to rely on a 50% chance level. Our simulations will also help researchers to determine whether it is better to increase the number of trials or rather whether it is better to increase the number of participants.

METHOD

Simulations

Analyses were run in R 3.3.2 (R Development Core Team, 2008) we have incorporated the R script as Electronic Supplementary Materials. We simulated sample sizes of 30 to 120 (with increments of 10) and for a range of trials (2,4,6,10,14,18,22,26,30,34). The probability is set at .5 for each trial. Needless to say that these are arbitrary choices and we offer our script should researchers want to test other values of the above. Researchers

can use the included script to calculate the probability under chance. The script simulates n binomial trials (where n is the range of trials) with the probability of .5. It does so 100,000 times for each of the given sample sizes (N). The analyses were run in duplicate (with two different random starting seeds). There was near perfect correspondence between the two runs (for all trials: all Pearson r>.99) we therefore averaged across both runs. We report the means, 2.5% and 97.5% percentiles of the simulations. These are labelled as 'true probability', the means should be around .5 but of interest are the confidence intervals (2.5 and 97.5 percentiles).

RESULTS

The results are summarized in Figure 1 (also see the appendix). As expected the means was always close to .5, which justifies the use of this cut-off. However, there was stark variation in the confidence intervals around the .5 cut-off. When using only two trials it appears that .5 is a weak benchmark, especially with smaller sample sizes. Even with larger sample sizes (n=120), it would be advisable to test against 55% or higher when relying on two trials. Increasing the number of trials allows setting a lower threshold for 'true probability' and as we increased the number of trials 50% becomes an increasingly acceptable benchmark. However, it is clear that the payoffs have diminishing returns, moving from 2 to 8 trials mattered more than moving from 26 to 34.

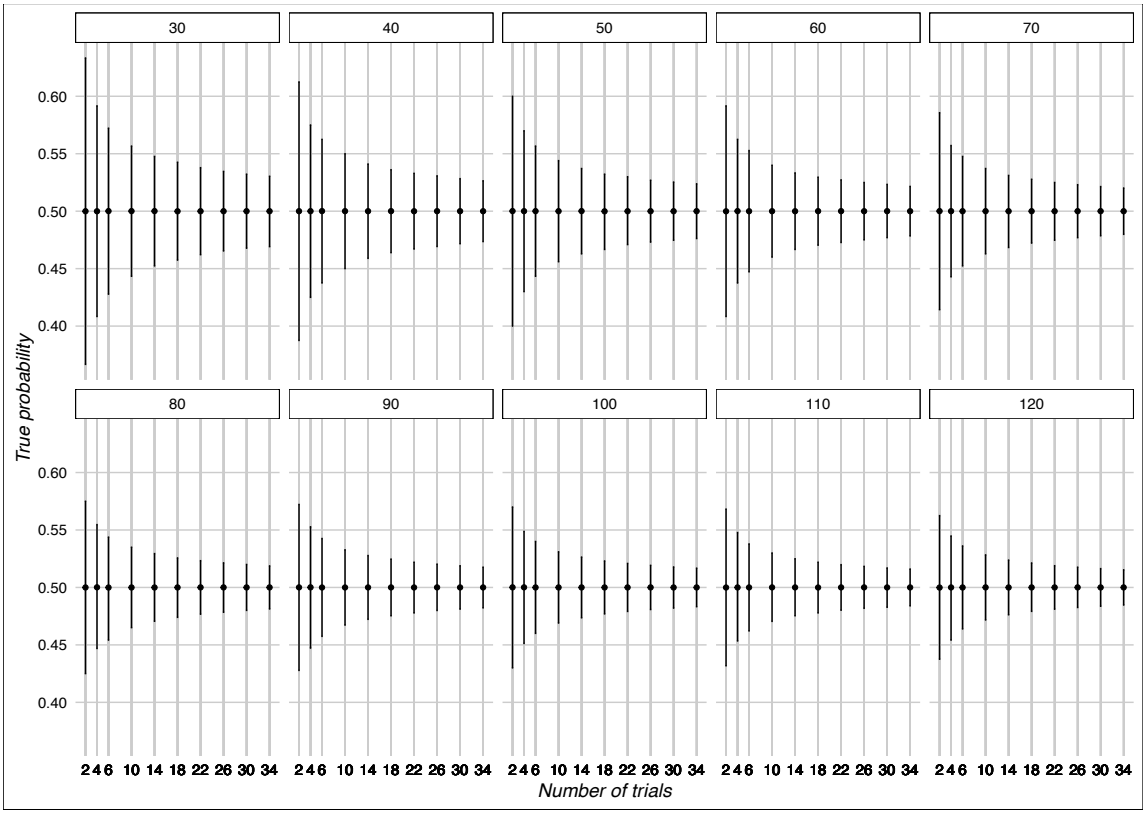

Figure 1: True' probability of binomial trials with 50% chance rate (means and 95%CI (2.5 and 97.5 percentile from simulations). The panels display various sample sizes, the X-axis indicates the number of trials.

The script we include allows researchers to plot their results against chance. As an illustration, we plot the result from Little et al. (2007) Study 1, where they report 57% of individuals voting for the "winning face" with a sample of 110 individuals and eight forced choices, against 100,000 simulations. As is clear from Figure 2, the result is upheld when applying this simulation approach: Individuals "voted" for the winning faces more often than the losing faces than expected under chance. In fact, only 1 out of 100,000 simulations scored 57% or higher.

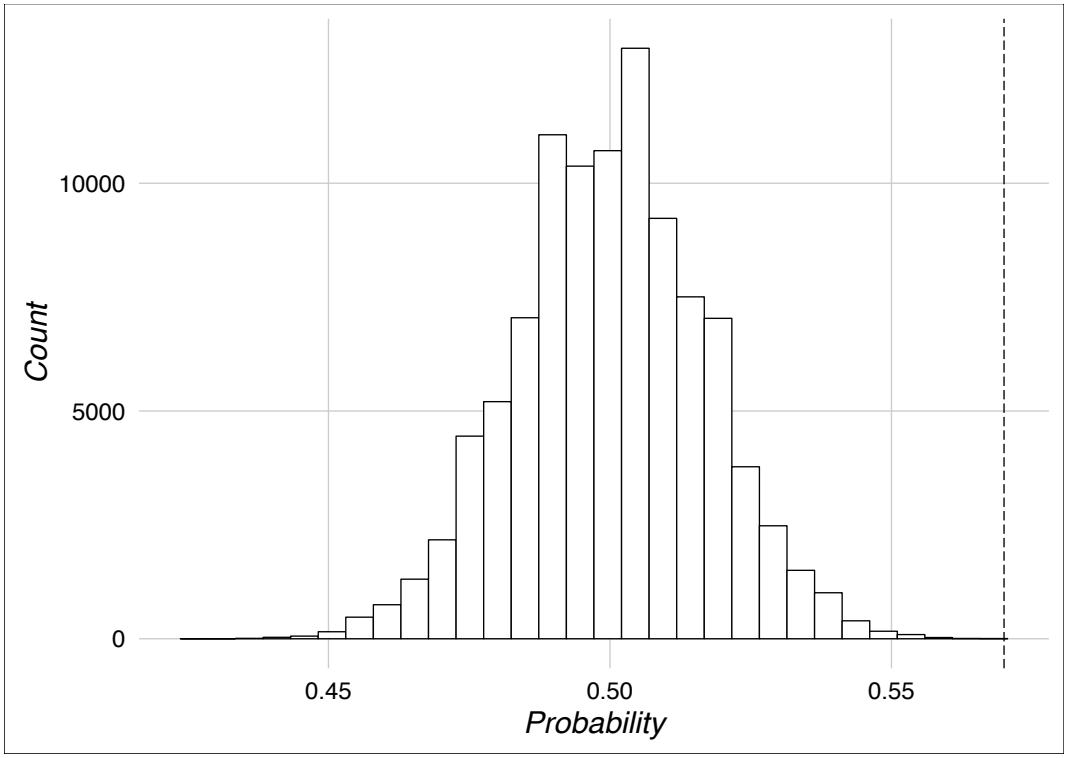

Figure 2: The probability of 100,000 simulations (n= 8 trials, N=110 participants, 50% chance) and the result from Little et al. 2007 study 1, (dashed line at 57% chance (0.57)).

DISCUSSION AND CONCLUSION

In this short methodological note, we examined if using 50% as a benchmark for chance is appropriate for forced choice tests with two alternatives at a range of trial numbers and sample sizes. Perhaps our results are unsurprising: the results indicate that with a small number of trials (n=2) using 50% is a rather weak threshold for evidence. If one wants to

have a better test that the results are not due to chance, then it is advisable to use a more conservative estimate, such as 55% or 60% when testing forced choice results that are thought to represent greater than chance, or increase the number of forced choice trials. Moreover, using the script we provide researchers can plot their result against a large number of simulations. Rather than testing against chance, researchers can thus simulate the likelihood of their findings compared to simulated data and we have included such an illustration.

We should reiterate that our paper does not deal with issues regarding to pooling across trials (e.g., Kievit et al., 2013; Pollet et al., 2015). When averaging across multiple trials, it is possible that a small number of trials are driving the effect. Apart from testing against a pooled estimate, it is therefore desirable to examine the results at trial level as well, or to employ a multilevel approach. In addition, we have not discussed issues with pseudo-replication which apply to these designs (e.g., Baayen, Davidson, & Bates, 2008; Waller, Warmelink, Liebal, Micheletta, & Slocombe, 2013). That being said, moving towards more conservative estimates could already be a useful first step for researchers working with forced choice data.

More broadly our paper demonstrates the usefulness of randomization and simulation approaches and as others have argued this forms a useful addition to the methodological toolbox (e.g., Stulp & Barrett, 2013). Finally, it should be noted that we have restricted ourselves to the 2AFC design but similar approaches can be developed for multi-choice designs or ranking methods.

In conclusion, we recommend that when researchers are using a small number of forced choice trials that they would modify the benchmark against which they test. If so desired researchers can actually simulate their experimental design and compare their data to the simulations. We believe that such simulations might provide a more rigorous test than the use of a 50% base line.

REFERENCES

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. DOI

Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, *26*(4), 313–331. DOI

Bendixen, M., Kennair, L. E. O., & Buss, D. M. (2015). Jealousy: Evidence of strong sex differences using both forced choice and continuous measure paradigms. *Personality and Individual Differences*, *86*, 212–216. DOI

Bleske-Rechek, A., Nelson, L. A., Baker, J. P., Remiker, M. W., & Brandt, S. J. (2010). Evolution and the trolley problem: People save five over one unless the one is young, genetically related, or a romantic partner. *Journal of Social, Evolutionary, and Cultural Psychology*, *4*(3), 115. DOI

Bressler, E. R., & Balshine, S. (2006). The influence of humor on desirability. *Evolution and Human Behavior*, *27*(1), 29–39. DOI

Buller, D. J. (2005). Evolutionary psychology: the emperor's new paradigm. *Trends in Cognitive Sciences, 9*(6), 277–283. DOI

Buss, D. M., Larsen, R. J., Westen, D., & Semmelroth, J. (1992). Sex Differences in Jealousy: Evolution, Physiology, and Psychology. *Psychological Science, 3*(4), 251–255.DOI

DeBruine, L. M. (2004). Facial resemblance increases the attractiveness of same–sex faces more than other–sex faces. *Proceedings of the Royal Society of London. Series B: Biological Sciences, 271*(1552), 2085 LP-2090. DOI

Haselton, M. G., & Miller, G. F. (2006). Women's fertility across the cycle increases the short-term attractiveness of creative intelligence. *Human Nature, 17*(1), 50–73. DOI

Kievit, R. A., Frankenhuis, W. E., Waldorp, L. J., & Borsboom, D. (2013). Simpson's paradox in psychological science: a practical guide. *Frontiers in Psychology, 4*, 513. DOI

Kurzban, R., DeScioli, P., & Fein, D. (2012). Hamilton vs. Kant: pitting adaptations for altruism against adaptations for moral judgment. *Evolution and Human Behavior, 33*(4), 323–333. DOI

Li, N. P., & Kenrick, D. T. (2006). Sex Similarities and Differences in Preferences for Short-Term Mates: What, Whether, and Why. *Journal of Personality and Social Psychology, 90*(3), 468–489. DOI

Little, A. C., Burriss, R. P., Jones, B. C., & Roberts, S. C. (2007). Facial appearance affects voting decisions. *Evolution and Human Behavior, 28*(1), 18–27. DOI

Macmillan, N. A. (2002). Stevens' Handbook of Experimental Psychology. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology*. Hoboken, NJ, USA: John Wiley & Sons, Inc. DOI

Penton-Voak, I. S., & Chen, J. Y. (2004). High salivary testosterone is linked to masculine male facial appearance in humans. *Evolution and Human Behavior, 25*(4), 229–241. DOI

Perrett, D. I., Lee, K. J., Penton-Voak, I., Rowland, D., Yoshikawa, S., Burt, D. M., … Akamatsu, S. (1998). Effects of sexual dimorphism on facial attractiveness. *Nature, 394*(6696), 884–887. DOI

Pollet, T. V., Stulp, G., Henzi, S. P., & Barrett, L. (2015). Taking the aggravation out of data aggregation: A conceptual guide to dealing with statistical issues related to the pooling of individual-level observational data. *American Journal of Primatology, 77*(7), 727–740. DOI

R Development Core Team. (2008). R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Rhodes, G., Proffitt, F., Grady, J. M., & Sumich, A. (1998). Facial symmetry and the perception of beauty. *Psychonomic Bulletin & Review, 5*(4), 659–669. DOI

Roberts, S. C., Havlicek, J., Flegr, J., Hruskova, M., Little, A. C., Jones, B. C., … Petrie, M. (2004). Female facial attractiveness increases during the fertile phase of the menstrual cycle. *Proceedings of the Royal Society B: Biological Sciences, 271*(Suppl_5), S270–S272. DOI

Schützwohl, A. (2004). Which infidelity type makes you more jealous? Decision strategies in a forced-choice between sexual and emotional infidelity. *Evolutionary Psychology, 2*(1), 121–128. DOI

Stirrat, M., & Perrett, D. I. (2010). Valid facial cues to cooperation and trust male facial width and trustworthiness. *Psychological Science, 21*(3), 349–354. DOI

Stulp, G., & Barrett, L. (2013). Binomial Tests and Randomization Approaches: The Case of US Presidential Candidate Height and Election Outcomes. *SAGE Research Methods Cases*. DOI

Tovée, M. J., Edmonds, L., & Vuong, Q. C. (2012). Categorical perception of human female physical attractiveness and health. *Evolution and Human Behavior*, *33*(2), 85–93. DOI

Verplaetse, J., Vanneste, S., & Braeckman, J. (2007). You can judge a book by its cover: the sequel.: A kernel of truth in predictive cheating detection. *Evolution and Human Behavior*, *28*(4), 260–271. DOI

Waller, B. M., Warmelink, L., Liebal, K., Micheletta, J., & Slocombe, K. E. (2013). Pseudoreplication: a widespread problem in primate communication research. *Animal Behaviour*, *86*(2), 483–488. DOI

ELECTRONIC SUPPLEMENTARY MATERIALS

- Reference table based on the simulations for various sample sizes, trials (means and 95%CI). see main text and Figure 1.

- R script.