A Preliminary P-Curve Meta-Analysis of Learned Categorical Perception Research

Janet Andrews (andrewsj@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Joshua de Leeuw (jdeleeuw@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Calais Larson (calarson@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Xiaoqing Xu (xixu@vassar.edu)
Department of Cognitive Science, 124 Raymond Ave.
Poughkeepsie, NY 12604 USA

Abstract

A preliminary meta-analysis using the p-curve method (Simonsohn, Nelson, & Simmons, 2014) was performed on a subset of the learned categorical perception literature to explore the robustness of the phenomenon. Only studies using novel visual categories and behavioral measures were included. The results strongly suggest that the phenomenon is robust but that the studies are somewhat underpowered. We argue that this is problematic because it renders both statistically significant and nonsignificant results very difficult to interpret, which impedes progress in understanding the learned CP phenomenon, for example, why expansion vs. compression is observed, or boundary vs. dimensional effects. Fortunately, there is a clear solution: conduct studies with greater statistical power.

Keywords: categorical perception; categorization; learning; p-curve; statistical power; expansion; compression; dimensional modulation

Introduction

Learned categorical perception (CP) is a phenomenon whereby learning to place objects into categories alters some aspect of the way those objects are judged (for a review, see Goldstone & Hendrickson, 2009). The classic patterns of change in judgments are that items placed in different categories become more distinguishable, sometimes called expansion, and/or that items placed in the same category become less distinguishable, sometimes called compression. These are category boundary effects, but other versions of learned CP are increased sensitivity to dimensions relevant to the category distinction and/or decreased sensitivity to dimensions irrelevant to the category distinction.

There has been a great deal of research on learned CP, but this paper will focus on visual categories learned in a laboratory setting. This is motivated mainly by our own interest in learned visual CP and the recognition that there may be important differences between CP in different modalities that would make it inappropriate to group those studies together for this meta-analysis. Thus the large body of research on auditory CP, in particular for speech sound distinctions that are acquired in the lengthy process of learning a natural language, will not be considered here. It is notable that laboratory-induced learned CP effects are obtained with very little training compared to the kind of exposure that is usually given in real category learning, e.g., learning color categories. This makes the phenomenon appear to be pervasive and basic, but there are several important questions that need to be addressed.

First, in light of the recent attention given to failures to replicate (Open Science Collaboration, 2015), p-hacking (Head, Holman, Lanfear, Kahn, & Jennions, 2015), the file drawer problem (Rosenthal, 1979), lack of sufficient statistical power (Button, Ioannidis, Mokrysz, Nosek, Flint, Robinson, & Munafò, 2013), and so forth, we believe that it is essential to assess whether the published learned CP literature demonstrates convincing evidence for the effect. Like many, perhaps most, areas of research in cognitive science, the phenomenon of learned CP allows researchers many degrees of freedom that may, unintentionally no doubt, inflate the appearance of real effects. For example, within a single study, there are often a variety of different ways in which the data can be analyzed to look for evidence of learned CP effects. Researchers can investigate accuracy and/or response time data – both have been used as evidence of CP in the literature – in a multitude of combinations due to the many different possible behavioral patterns that count as learned CP. Furthermore, different potential criteria for what counts as successful learning (a precondition for testing for learned CP effects) can be used, leading to additional choices that can unintentionally bias the analysis. In short, this is exactly the kind of situation where preregistration is important to avoid mistaking the noise in the data for signal. To our knowledge, very little if any learned CP research has been preregistered as of yet, for replication purposes or otherwise.

One reason this is important is the current controversy over whether there are really any genuine top-down effects of cognition on perception, of which learned CP could potentially be one type. Firestone and Scholl (2016) argue that none of the vast amount of research claiming to have demonstrated such effects has actually successfully done so. It seems to us that before we can effectively debate whether learned CP provides evidence for such effects that is immune from Firestone and Scholl's criticisms, we must first establish that there is credible evidence that the effects themselves actually exist.

The optimist might point to the dozens of studies on CP that report significant results as compelling evidence that these effects are real. The problem with this approach is that it ignores the existence of publication bias in favor of statistically significant results. If learned CP turned out to not be replicable, it would not be the first example of a widely reported phenomenon that did not reliably replicate (e.g., Lurquin et al., 2016; Papesh, 2015; Shanks et al., 2013; Simmons & Simonsohn, in press).

The purpose of this paper was to do a preliminary evaluation of published learned CP effects by using a p-curve analysis (Simonsohn, Nelson, & Simmons, 2014). A p-curve is a meta-analytic technique that looks at the distribution of statistically significant p-values in a set of related studies. An advantage of the p-curve approach is that it nicely handles the file-drawer problem by looking only at statistically significant p-values. If the results come from a collection of well-powered studies investigating a real effect, most of the p-values should be very small (well below .05). However, if many of the results are due to false-positives, either because of a lack of statistical power or because the phenomenon being studied is not real, then the distribution of significant p-values will be flat, in the case of a null effect, or close to flat, in the case of low statistical power. (In extreme cases, the presence of substantial p-hacking can generate a p-curve with left skew.) Furthermore, the p-curve can reliably estimate the average statistical power of a set of studies because the distribution of observed p-values is directly related to statistical power when the null hypothesis is false. This estimation of power is an average estimate assuming that all studies are investigating the same basic effect.

In the analyses presented below, we calculate p-curves for a set of studies from the visual learned CP literature, but we caution that this analysis is preliminary. We expect that enlarging the scope of the search for relevant sources would yield many additional studies of learned CP that could be included in the analysis, though our sample size is large enough to likely be informative.

Method and Results

The articles used in the analysis were selected by conducting a Scopus search of all articles citing Goldstone (1994) that had "CP" or "categorical perception" in the title, abstract, or key words and were deemed relevant (i.e., reported original empirical results in detail; used visual

Table 1. Articles used in the p-curve meta-analysis of learned categorical perception research.

Authors	Year of publication	# experiments included	# analyses included	Learned CP measure	Type of measure[a]
Corneille & Judd	1999	3	2	Typicality	S
Folstein, Palmeri, & Gauthier	2014	1	1	Same-different	O
Goldstone	1994	4	8	Same-different	O
Goldstone, Lippa, & Shiffrin	2001	1	1	Similarity	S
Goldstone, Steyvers, & Larimer	1996	1	1	Same-different	O
Grandison, Sowden, Drivonikou, Notman, Alexander, & Davies	2016	1	1	Target location RT	O
Gureckis & Goldstone	2008	1	1	XAB	O
Holmes & Wolff	2012	1	1	Discrimination RT	O
Levin & Beale	2000	3	2	XAB variant	O
Livingston & Andrews	2005	2	2	Similarity, same-different	S, O
Livingston, Andrews, & Harnad	1998	3	2	Similarity	S
Notman, Sowden, & Özgen	2005	2	2	Same-different	O
Op de Beeck, Wagemans, & Vogels	2003	2	2	Same-different	O
Özgen & Davies	2002	2	2	Same-different	O
Stevenage	1998	2	4	Similarity	S
Zhou, Mo, Kay, Kwok, Ip, & Tan	2010	1	1	Target location RT	O

[a]O = objective, S = subjective

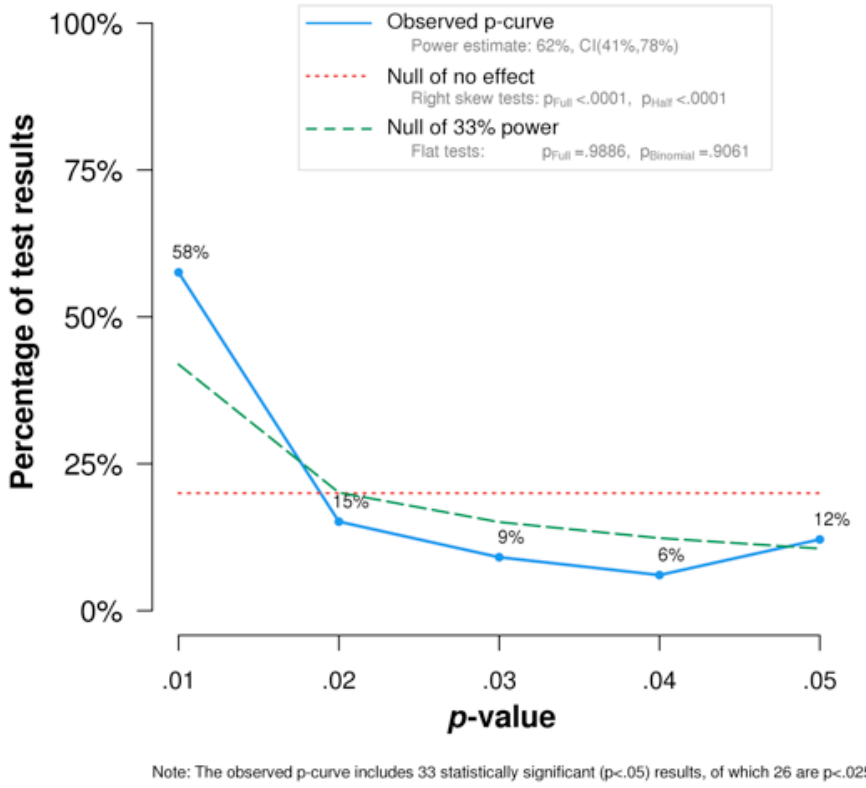

Figure 1: P-curve for all relevant results in the articles listed in Table 1. The solid blue line shows the observed distribution of significant p-values. The dotted red line shows what the expected distribution of p-values would be if the null hypothesis were true. The right skew tests, for both the full set of p-values (all p-values < .05) and half set of p-values (all p-values < .025), indicate that the p-values are more right skewed than would be expected if the null were true. The dashed green line represents the expected distribution of p-values if the set of studies had 33% power. The flat tests test if the observed distribution is flatter than the distribution that would be observed if the studies had 33% power.

stimuli, unfamiliar categories, and behavioral measures; and focused on learned CP). This yielded 14 articles; in addition, we included two conference papers meeting the same criteria for a total of 16 sources (see Table 1). Within those 16 sources were a total of 30 experiments reporting 42 distinct relevant statistical results. (For example, statistical results pertaining to the learning of the categories per se were not relevant.) Of these 42 statistical results, 33 were statistically significant in the predicted direction and these results were input to the p-curve app version 4.05 (http://www.p-curve.com/app4/) to produce the p-curve shown in Figure 1. Note that the p-curve analysis only considers the distribution of p-values below the 0.05 threshold.

In addition, separate p-curves were generated for the subset of results obtained using objective measures of learned CP, such as accuracy of same-different judgments, and the subset of results obtained using subjective measures such as similarity judgments. These are shown in Figures 2 and 3.

The p-curves shown in Figures 1-3 display the distribution of p-values that fall into five bins. For a real effect with high power samples, most of the p-values should be in the leftmost bin ($p < 0.01$). Shown in the figures are what the distribution would look like with a set of studies powered at 33% (green dashed line) and a set of studies testing a null effect (red dotted line). Note that because the distribution of p-values is determined by statistical power when investigating a real effect, the average power of the studies can be estimated from the observed distribution of p-values.

The results of the overall p-curve meta-analysis show a curve that is right-skewed, which strongly suggests that the research has evidential value and is not the result of worrisome p-hacking (which produces a left-skewed curve).[1] This is welcome news. However, the estimated power is only 62% (90% confidence interval is 41-78%) which is not very high. If this estimate were correct, it would mean that only 6 in 10 studies of learned CP will detect an effect. We explain why this is a significant

[1]The p-curve app conducts three different statistical tests to detect whether there is right skew. All of these tests were statistically significant ($p < 0.001$ for all), indicating that the curve is very unlikely to be not right skewed.

problem for advancing our theory of learned CP in the discussion.

We conducted a separate analysis for objective and subjective measures of CP for two reasons. The first is that some researchers have suggested based on neural evidence that learned CP is not a genuinely perceptual effect but occurs at a higher, post-perceptual level (e.g., Clifford, Franklin, Holmes, Drivonikou, Özgen, & Davies, 2012; but see Zhong, Li, Li, Xu, & Mo, 2015 for counterevidence). We reasoned that if this were the case, studies employing more subjective measures of learned CP, such as ratings of similarity or typicality, should show stronger effects. The second is that there is a worry that learned CP effects could be the result of demand effects (Goldstone, Lippa, & Shiffrin, 2001). That is, participants in these experiments may indicate that two items are subjectively more similar (dissimilar) to each other precisely because the experiment trains them that the two objects belong to the same category (different categories), and not because of any perceived change in similarity of the visual objects. If this is a contributing factor, then we would expect studies with subjective measures to have higher power, because this demand effect will only contribute for subjective measures of CP.

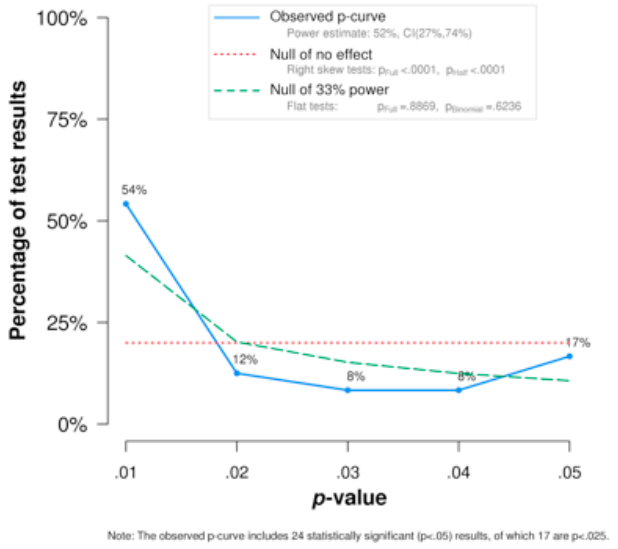

Figure 2: P-curve for learned CP results based on objective measures.

These ideas received some support from the p-curve patterns based on studies using subjective vs. objective learned CP measures, with a generally stronger pattern and higher power estimate for the subjective measure studies (79%) than objective measure studies (52%). However, given the preliminary nature of this analysis and the small set of results included, particularly for those using subjective measures, it is premature to draw any conclusions about this yet (note that the confidence intervals for the power estimates overlap substantially). Furthermore, we can't distinguish between the two possible explanations of this result without directly investigating the matter.

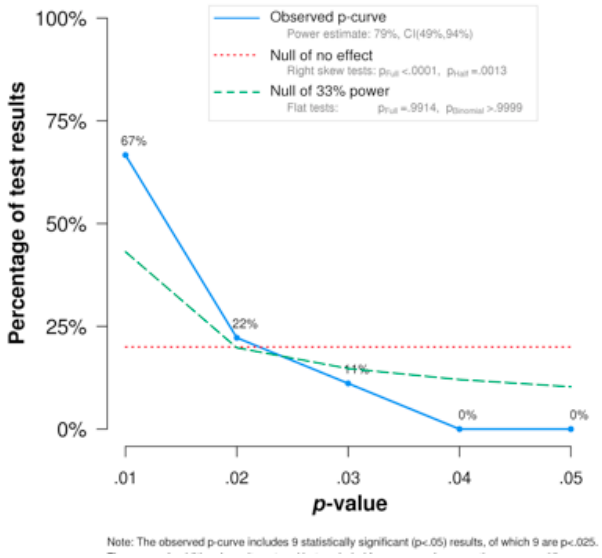

Figure 3: P-curve for learned CP results based on subjective measures.

Discussion

The literature for learned visual categorical perception contains evidentiary value, according to this meta-analysis. We can be reasonably confident that the studies reported are in general not reporting on a null effect. However, the relatively low statistical power shown by this analysis for the overall set of findings has important implications for how our theoretical understanding of learned CP is informed by these studies, and future studies with similar statistical power. We argue here that the statistical power of learned CP research must be improved in order to make robust advances in theory.

Several debates in learned CP research (e.g., to what degree is CP a perceptual or decision-making process; what kinds of judgments are changed by learning categories; is CP the result of demand effects) currently hinge on the observation of CP in some experimental contexts but not others. However, the overall lack of statistical power makes the pattern of significant and non-significant results difficult to interpret. Low power may well explain the occurrence of non-significant results. Low power also increases the likelihood that significant results are actually false positives. It follows from these two facts that when studies of learned CP are underpowered, the noise in the data makes it very difficult to distinguish among specific theoretical variants of what learned CP is. For example, it is difficult to distinguish among the different types of learned CP, of which there are at least four, as noted in the introduction (the boundary effects of compression and/or expansion and dimensional sensitization and desensitization based on category relevance). Since null results are impossible to interpret when statistical power is low (and using traditional

statistical methods), and patterns in statistically significant results may be just noise, it is correspondingly impossible to use the data to figure out under what conditions each of the types of learned CP do and do not occur. Yet this is essential to do in order to determine the nature of learned CP mechanisms and their purpose.

To understand the problem that this causes for our theoretical understanding of learned CP, it is important to keep in mind that the articles included in our analysis are not a set of direct replications, but rather a body of scientific evidence. Experiments in this set of results aim to build on the contributions of prior work to refine our theory of learned CP. Thus, we often rely on the pattern of findings within individual studies, or between small sets of studies, to constrain theorizing. But, as noted, with low statistical power comes the increased probability of false negatives and the increased probability that significant results are false positives. This leaves the theorist in a tough position. Are we improving our theoretical understanding with a new set of data, or merely reading the tea leaves of statistical noise?

Recent work in our lab provides an illustration of this problem (de Leeuw, Andrews, Livingston, & Chin, 2016). We were primarily interested in why some learned CP studies had shown compression while others showed expansion. There seemed to be a relationship between both the type of learned CP measure (similarity vs. same-different judgment accuracy) and stimulus discriminability, on the one hand, and the pattern of learned CP on the other. Initial studies in our lab seemed to confirm variations of this kind but the patterns were somewhat bewildering. Only when we conducted a large scale study (N > 550) simultaneously incorporating multiple measures and levels of stimulus discriminability and used Bayesian data analysis did a clear picture emerge: learned CP effects occurred (in fact, three of the four possible patterns occurred), but this pattern of effects did not differ systematically according to either of those variables.[2]

It is important to note a related set of problems with learned CP research that also presented a challenge for conducting the p-curve analysis. (1) Learned CP studies often don't test for more than one or two of the four possible types of effect, and the statistical analysis used may test for different effects separately or lump them together. Furthermore, there are methodological ambiguities in many of the studies that make separating out which effects occurred impossible.[3] We therefore could not classify the p-values according to which aspect of learned CP they corresponded to, even though we would have liked to be able to do this. (2) Predictions are often vague in regard to the nature of a two-way interaction in an ANOVA, for example. But different results should be used for the p-curve for attenuation and reversal interaction predictions (just the overall interaction for attenuation and just the simple effects for reversal). Since we were limited to the information available in the articles, most of which did not predict a specific pattern of interaction, power may be overestimated by the p-curve. (3) A final caveat regarding our analysis is that our p-curve results could potentially be somewhat misleading if in fact certain learned CP effects (e.g., dimensional effects) are much stronger than others (e.g., boundary effects), which would mean that they are not really the same kind of effect as assumed by a combined p-curve analysis. If this were the case, it would suggest that power could be higher than our estimate for some aspects of learned CP, but lower for others. This would only further exacerbate the problem of drawing theoretical conclusions about the nature of learned CP, as the studies of certain kinds of effects that are deeply relevant to the theory would have even lower power.

The preliminary meta-analysis we report here strongly suggests that learned CP effects are real but also that our current knowledge of them is highly ambiguous and destined to remain so if we do not change the way we do research. In our view, only by conducting future studies with sufficient statistical power will we make significant progress understanding the phenomenon of learned CP.

[2] This study was not included in the preliminary meta-analysis reported here because it did not provide the relevant standard statistical information.

[3] Another difficulty with sorting out the different potential effects of learned CP is that nearly all the published work is not designed to address this question. This makes it impossible in most cases to distinguish between, as an example, acquired distinctiveness of a dimension plus compression versus just expansion. These two cases have the same behavioral outcome, and thus must be distinguished through experimental controls, such as different kinds of training.

References

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365-376.

Clifford, A., Franklin, A., Holmes, A., Drivonikou, V. G., Özgen, E., & Davies, I. R. L. (2012). Neural correlates of acquired color category effects. *Brain and Cognition, 80*(1), 126-143.

Corneille, O., & Judd, C. M. (1999). Accentuation and sensitization effects in the categorization of multifaceted stimuli. *Journal of Personality and Social Psychology, 77*(5), 927-941.

de Leeuw, J. R., Andrews, J. K., Livingston, K. R., & Chin, B. M. (2016). The effects of categorization on perceptual judgment are robust across different assessment tasks. *Collabra, 2*(1), 1–9.

Firestone, C., & Scholl,, B. (2016). Cognition does not affect perception: Evaluating the evidence for "top-down" effects. *Behavioral and Brain Sciences, 39*.

Folstein, J. R., Palmeri, T. J., & Gauthier, I. (2014). Perceptual advantage for category-relevant perceptual dimensions: The case of shape and motion. *Frontiers in Psychology, 5*.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*, 178-200.

Goldstone, R. L., & Hendrickson, A. T. (2009). Categorical perception. *Interdisciplinary Reviews: Cognitive Science*, *1*, 69–78.

Goldstone, R. L., Lippa, Y., & Shiffrin, R. M. (2001). Altering object representations through category learning. *Cognition, 78*(1), 27-43.

Goldstone, R. L., Steyvers, M., & Larimer, K. (1996). Categorical perception of novel dimensions. *Proceedings of the eighteenth annual conference of the cognitive science society* (pp. 243-248). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grandison, A., Sowden, P. T., Drivonikou, V. G., Notman, L. A., Alexander, I., & Davies, I. R. L. (2016). Chromatic perceptual learning but no category effects without linguistic input. *Frontiers in Psychology, 7*.

Gureckis, T. M., & Goldstone, R. L. (2008). The effect of internal structure of categories on perception. *Proceedings of the thirtieth annual conference of the cognitive science society* (pp. 1876-1881). Washington, D.C.: Cognitive Science Society.

Head, M. L., Holman L., Lanfear, R., Kahn A. T., & Jennions M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106.

Holmes, K. J., & Wolff, P. (2012). Does categorical perception in the left hemisphere depend on language? *Journal of Experimental Psychology: General, 141*(3), 439-443.

Levin, D. T., & Beale, J. M. (2000). Categorical perception occurs in newly learned faces, other-race faces, and inverted faces. *Perception and Psychophysics, 62*(2), 386-401.

Livingston, K. R., & Andrews, J. K. (2005). Evidence for an age-independent process in category learning. *Developmental Science, 8*(4), 319-325.

Livingston, K. R., Andrews, J. K., & Harnad, S. (1998). Categorical perception effects induced by category learning. *Journal of Experimental Psychology: Learning Memory and Cognition, 24*(3), 732-753.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., & Miyake, A. (2016). No evidence of the ego-depletion effect across task characteristics and individual differences: A pre-registered study. *PLoS ONE, 11*(2), e0147770.

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: A psychophysical approach. *Cognition, 95*(2), B1-B14.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2003). The effect of category learning on the representation of shape: Dimensions can be biased but not differentiated. *Journal of Experimental Psychology: General, 132*(4), 491-511.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Özgen, E., & Davies, I. R. L. (2002). Acquisition of categorical color perception: A perceptual learning approach to the linguistic relativity hypothesis. *Journal of Experimental Psychology: General, 131*(4), 477-493.

Papesh, M. (2015). Just out of reach: On the reliability of the action-sentence compatibility effect. *Journal of Experimental Psychology: General, 144*(6), e116-e141.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*(3), 638-641.

Shanks, D.R., Newell, B.R., Lee, E.H., Balakrishnan, D., Ekelund, L., Cenac, Z., et al. (2013). Priming intelligent behavior: an elusive phenomenon. *PLoS ONE*, *8*(4), e56515.

Simmons, J., & Simonsohn, U. (in press). Power posing: P-curving the evidence *Psychological Science*.

Simonsohn, U., Nelson, L., & Simmons, J. (2014). P-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547.

Stevenage, S. V. (1998). Which twin are you? A demonstration of induced categorical perception of identical twin faces. *British Journal of Psychology, 89*(1), 39-57.

Zhong, W., Li, Y., Li, P., Xu, G., & Mo, L. (2015). Short-term trained lexical categories produce preattentive categorical perception of color: Evidence from ERPs. *Psychophysiology, 52*(1), 98-106

Zhou, K., Mo, L., Kay, P., Kwok, V. P. Y., Ip, T. N. M., & Tan, L. H. (2010). Newly trained lexical categories produce lateralized categorical perception of color. *Proceedings of the National Academy of Sciences of the United States of America, 107*(22), 9974-9978.