Bilingualism: Language and Cognition

cambridge.org/bil

Power considerations in bilingualism research: Time to step up our game

Marc Brysbaert

Ghent University, Belgium

Review Article

Cite this article: Brysbaert M (2020). Power considerations in bilingualism research: Time to step up our game. *Bilingualism: Language and Cognition* 1–6. https://doi.org/10.1017/S1366728920000437

Received: 22 January 2020
Revised: 22 June 2020
Accepted: 22 July 2020

Keywords:
Bilingualism; power; replication

Address for correspondence:
Marc Brysbaert,
E-mail: marc.brysbaert@ugent.be

Abstract

Low power in empirical studies can be compared to blurred vision. It makes the signal ambiguous, so that conclusions depend more on interpretation than on observation. Data patterns that look sensible are published as evidence for theoretical positions and unclear patterns are discarded as noise, whereas both could be due to sampling error or could be a perfect reflection of the population parameters. Simulations indicate that little research with sample sizes lower than 100 participants per group provides a picture of enough resolution to draw firm conclusions. This is particularly true for research comparing groups of people and involving interaction effects. As a result, it is to be feared that many findings in bilingualism research do not have a firm base, certainly not if they go beyond a simple comparison of two within-participants conditions.

Low power results in an ambiguous picture stressing interpretation over observation

Vision is blurred when the eyes do not refract light rays properly on the retina. Everyone wearing glasses or lenses knows this and it is often seen most spectacularly in toddlers. Before their vision deficiency is detected, they look clumsy and less smart than other kids. Then suddenly everything changes when they get proper vision, and they rapidly become attached to their glasses.

Low statistical power is like blurred vision and it is astonishing that researchers would actively opt for such a condition (depicted in Figure 1). It makes the evidence ambiguous so that extra interpretation is needed (a.k.a. educated guessing). Still, that is what bilingualism researchers have been doing for the past 50 years. We are deliberately looking at the world around us with unfocused lenses, constantly shouting to each other that there might be something significant out there without being able to have a proper look. We even developed a very sophisticated statistical machinery to extract the most out of blurred images.

Low power is not like jaywalking

For a long time, researchers have known about the low power of their experiments, but they thought it was a minor offence, a bit like jaywalking (Simmons, Nelson & Simonsohn, 2018). Their thoughts are something like: "I know it is not law-abiding, but there is no harm in it. The only person who can get hurt is me, when I fail to obtain the predicted, statistically significant effect."

We now know that the consequences are far more serious. First, a significant effect is more likely to be a false positive finding (reflecting the null hypothesis H_0) than a true positive finding (reflecting the alternative hypothesis H_1) when the power of the study is low (e.g., LeBel, Campbell & Loving, 2017). This can be illustrated with the following numerical example. Suppose H_0 is 10 times more likely than H_1 (a reasonable assumption if we are doing cutting edge research)[1]. Further suppose we use alpha = .05 and power = .80. Then suppose we run 1,100 studies. ||In 1,000 studies H_0 applies. Of these, 50 will be significant if we use $p < .05$. In 100 studies, H_1 applies and we will obtain a significant effect in 80 of them (due to our power). So, when we obtain a significant effect, chances of it pointing to a true finding (H_1) are 80/(50+80) = 62%. Now suppose, we run the same 1,100 studies with power = .40. Then we still have 50 significant false positives when H_0 applies, but we have only 40 significant effects when H_1 is valid. In other words, when we obtain a significant effect, chances that it reflects H_0 [50/(50+40) = 56%] are larger than chances that it reflects H_1 [40/(50+40) = 44%]. So, a significant effect is more likely to be a false positive than a true positive.

Second, when the outcome is not statistically significant, researchers have an incentive to "improve things", by running extra participants, by trying extra analyses, by excluding data

[1] Incidentally, this element also illustrates the importance of embedding research within a coherent theory. If a prediction is part of a well-substantiated theory, chances of H_1 being true can be made much higher than when the research is mostly a fishing expedition, even when the hypothesis is cutting edge (Loiselle & Ramchandra, 2015).

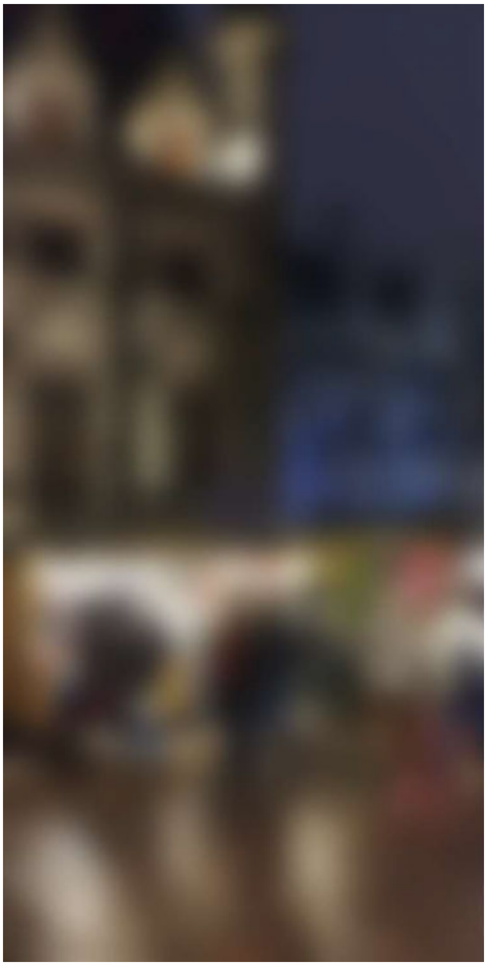


Fig. 1. Low statistical power is like blurred vision. It prevents us from having detailed information. What is depicted here? (See the Appendix for a clearer picture).

from bad participants, or by amending their hypotheses after the results are obtained (Gelman & Loken, 2013; Kerr, 1998; Simmons, Nelson & Simonsohn, 2011). These efforts increase the chances of finding statistical significance when there is none. As a result, they have been called questionable research practices (John, Loewenstein & Prelec, 2012) and they are known to contribute to the so-called replication crisis: the observation that fewer published findings are replicated than expected on the basis of statistical considerations (Maxwell, Lau & Howard, 2015; McElreath & Smaldino, 2015; Shrout & Rodgers, 2018).

Third, findings that fail to reach statistical significance are less likely to be published than significant findings, leading to the so-called file drawer problem (Rosenthal, 1979) and a biased literature (Brunner & Schimmack, 2020; De Bruin, Treccani & Della Sala, 2015). This is particularly true when data come from a study with low power. In case of a null effect, the spontaneous (and correct) reaction of most researchers is that "nothing can be concluded". In contrast, when a significant finding is obtained in a low power study, researchers (wrongly) assume that they have come across a big effect (otherwise it would not have been significant as shown in the figures below) and hence a "potentially important finding", worthwhile to be shared with the research community (Vasishth, Mertzen, Jäger & Gelman, 2018). The same is true for underpowered interactions: significant interactions that make sense (i.e., confirm our beliefs) are published, and the others are discarded.

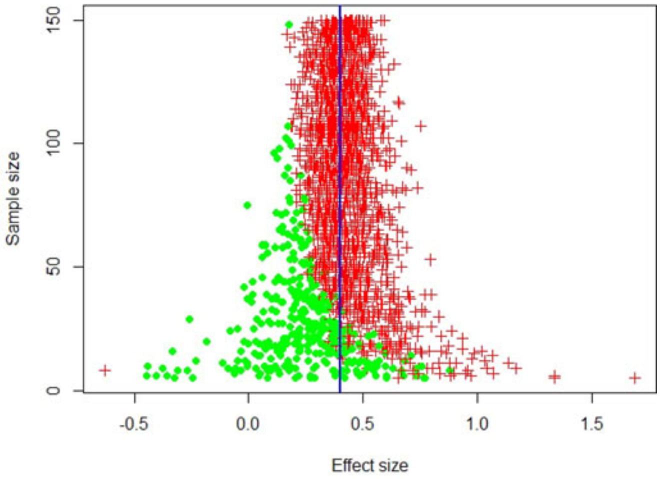


Fig. 2. Outcome of 2000 experiments trying to estimate a typical standardized effect size in psychology ($d = .4$, indicated by the blue vertical line) in a within-participants design with two conditions and participants taking part in both. Each symbol describes the outcome of a single study. It shows how large the effect was in the study and how many participants took part (ranging from 5 to 150). + signs indicate experiments with a statistically significant finding ($p < .05$); o signs indicate experiments that failed to reach significance.

All in all, deliberately running underpowered studies and pushing to get significant effects published not only leads to a blurred, ambiguous picture (Figure 1), but is wrong in a way that swindling is wrong (as opposed to jaywalking).

What is good power? Repeated measures

Figure 2 shows the outcome of a simulation with 2000 studies to test a typical effect size ($d = .4$; Brysbaert, 2019)[2] between two conditions in a repeated measures design (for instance, bilingual participants doing a task in their first and second language). The R commands for the simulation (and those for the other figures) are explained in detail in the supplementary materials, so that interested readers can adapt them to their needs. The numbers of participants per study differ from 5 to 150. Looking at the figure, most people would agree that a sample of 100-120 participants is a decent target: it allows you to get a pretty good estimate of the effect size in each study and the effect is always statistically significant. In contrast, sample sizes of less than 30 give a very blurry picture. You get divergent estimates of the effect size in different studies, and the effect size is seriously overestimated when you find statistical significance. As it happens, the correct conclusion (that there is a significant effect of $d = .4$) is almost never obtained in an experiment with fewer than 30 participants. So, running such a study more often increases the ambiguity in the literature instead of decreasing it. Remember that, if you run only one study, you do not have the advantage of the

[2] As argued by a reviewer, bilingual researchers may often be looking at effect sizes smaller than $d = .4$ (or $r = .2$), because variability tends to be larger in bilingual than in monolingual populations. The R programs referred to in the present article can easily be adapted for smaller effect sizes. Embrace yourself for brutal figures, however! Also ask yourself whether such research is still worthwhile, given that small effects only have practical consequences when they happen frequently and add up (Funder & Ozer, 2019). You need strong theoretical motivation to look for small effects.

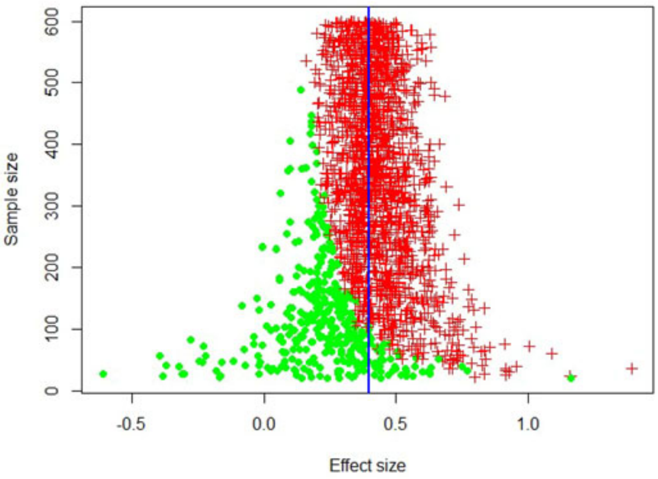


Fig. 3. Outcome of 2000 experiments trying to estimate a typical standardized effect size in psychology (d = .4, indicated by the vertical blue line) in a between-groups design with two conditions. Each symbol describes the outcome of a single study. It shows how large the effect was in the study and how many participants took part in the study (ranging from 20 to 600; 10-300 per condition). + signs indicate experiments with a statistically significant finding (p < .05); o signs indicate experiments that failed to reach significance.

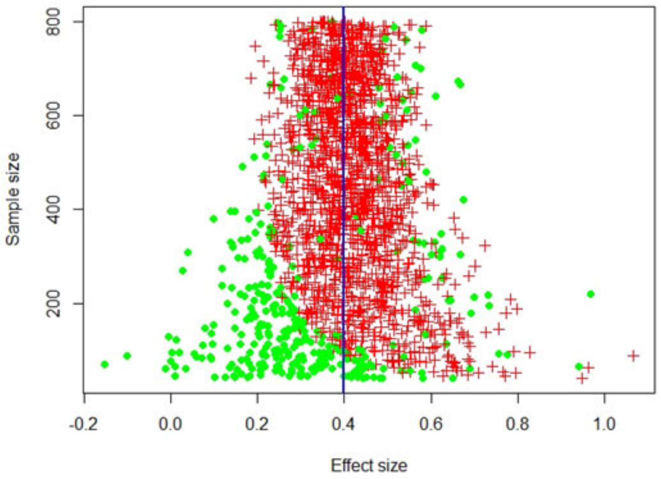


Fig. 4. Outcome of 2000 experiments trying to estimate the outcome of a 2×2 split-plot design, in which group 1 has an effect size of d = .4 and group 2 has an effect size of d = .0. Each symbol describes the outcome of a single study. It shows how large the interaction effect was in the study and how many participants took part in the study (ranging from 40 to 800; 20 to 400 per condition). + signs indicate experiments with a statistically significant interaction (p < .05), a statistically significant main effect for group 1 (p < .05) and no significant effect for group 2 (p > .05); o signs indicate experiments that failed to reach the pattern (for large sample sizes mostly because the effect in group 2 was significant). Remember that if you run only one study, you have only one data point and nothing to compare with. This is particularly a problem for sample sizes below 100 (50 per group), where the results cover all the range from -0.2 to +1.2 and rarely include the right conclusion (a significant interaction effect because one group has an effect of d = .4). Unfortunately, small sample sizes are the default option in the vast majority of studies on bilingualism.

bird's eye view shown in Figure 2. All you have is a single data point that can range from d < -.5 to d > 1.0 and is or is not statistically significant. On the basis of this single data point you draw theoretical conclusions. Also notice the small-size study with a significant effect in the OPPOSITE direction (entirely due to sampling error), which could seriously complicate the literature if published (Gelman & Carlin, 2014).

Unfortunately, many researchers do not have Figure 2 in mind when they design a study (also known as a funnel plot; Sterne, Becker & Egger, 2005). All they think of is the amount of work required for their study and how to get away with the smallest possible sample size (building on a tradition of similar sizes). As a result, sample sizes are more often closer to 20 than to 100.

Designs involving a between-groups variable require more participants, also for interactions with a repeated measure

Researchers on bilingualism have the extra complication that they often want to compare two groups of people: bilinguals versus monolinguals, or bilinguals with different degrees of proficiency. For instance, many articles have been published on the question of bilinguals having better executive control than monolinguals (for recent reviews, see Lehtonen, Soveri, Laine, Järvenpää, De Bruin & Antfolk, 2018; Paap, Mason, Zimiga, Silva & Frost, 2020). As is generally known, research between groups requires more participants. Figure 3 gives the same information as Figure 2, but now in a design that compares two groups of people. For such research, we easily need 300+ participants (150 per group) if we want to get a stable, clear picture. Notice how bad the situation is for sample sizes smaller than 100 (50 per group)! Still, of the 1004 studies reviewed by Lehtonen et al. (2018) 878 had sample sizes smaller than 50 participants per group (i.e., 87%) and 987 had sample sizes smaller than 100 (98%).

In large-scale replication attempts it has been found that in particular between-groups manipulations are difficult to replicate. This is understandable given the large sample sizes needed for unambiguous evidence (Figure 3).

What is generally not known is that Figure 3 also applies to a design in which a within-participants effect is compared across two groups, a so-called split-plot design. For instance, Kim (2020) compared Spanish heritage speakers with Spanish monolinguals on the processing of Spanish words that differed on the position of the lexical stress (penultimate or final syllable). In such a 2×2 design, the interaction has similar power requirements as the main effect of the between-groups variable (speaker group). This can be understood if you know that the interaction effect boils down to a between-groups t-test of the difference scores (e.g., Judd, McClelland, & Ryan, 2008). Try it out. Take the differences scores between the two within-participants conditions per participant (e.g., responses to words with final stress minus responses to words with penultimate stress) and run a between-groups one-way ANOVA on the difference scores. You will get the same F-value as the F-value of the interaction in the 2×2 design.

Because interactions between repeated measures and between-groups variables resemble comparisons between independent groups, it is to be feared that they will fare badly in replication attempts too, which is bad news for bilingualism research. As it happens, the situation is even more demanding than in Figure 3: because we not only want significant interactions, but interactions that agree with the model underlying the analysis. So, if the effect size for group 1 is d = .4 and the effect size for group 2 is d = .0, we not only want a significant interaction, but also a significant difference in the pairwise comparison for group 1 and no significant difference for group 2.

Figure 4 shows how often we obtain the required pattern as a function of the number of participants tested. As expected, it

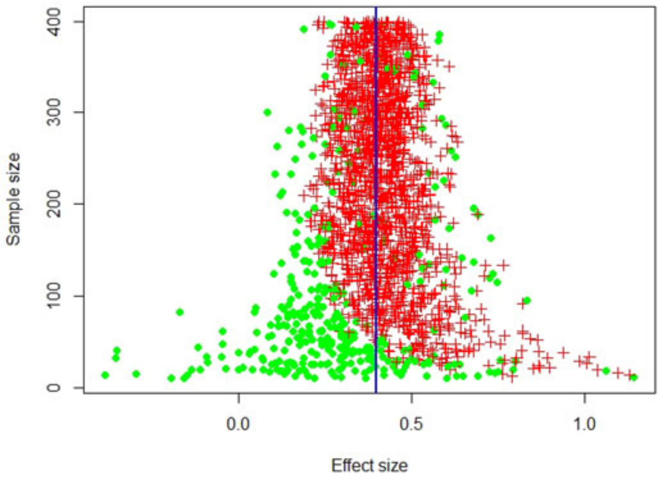


Fig. 5. Outcome of 2000 experiments trying to estimate the outcome of a 2×2 repeated measures design, in which variable B has an effect size of d = .4 at the first level of variable A and no effect size at the second level of A. Each symbol describes the outcome of a single study. It shows how large the interaction effect was in the study and how many participants took part in the study (ranging from 10 to 400). + signs indicate experiments with a statistically significant interaction ($p < .05$), a statistically significant main effect for variable B at the first level of variable A ($p < .05$) and no significant effect of variable B at the second level of variable A ($p > .05$); o signs indicate experiments that failed to reach the pattern. Remember that if you run only one study, you have but one data point and nothing to compare with.

looks much more like Figure 3 (between-groups effect) than like Figure 2 (within-participants design). The situation is even slightly worse, because we have studies without the full pattern for large numbers of participants, in line with the fact that interactions (involving a comparison of two difference scores) include more noise than main effects (involving only one difference score). It may be worthwhile to stress that the lowest sample size (the worst) already includes 40 participants; that is 20 per condition!

You also need more observations for interactions of within-participant variables

Also fully within-participant designs require more observations for interactions than for main effects (although thankfully not as many as an interaction with a between-groups variable). The effect size of an interaction is only as big as that of a main effect when the interaction is fully crossed: so, for an interaction of d = .4 in a 2×2 repeated measures design, you need d = +.4 for variable B at one level of variable A and d = -.4 at the other level. This pattern is virtually never expected. What is more likely is an effect of d = .4 at one level of variable A and no effect at the other level. This, however, effectively halves the effect size, meaning that you need four times as many participants (Brysbaert, 2019; Perugini, Gallucci & Costantini, 2018; Simonsohn, 2014). Furthermore, we not only want a significant interaction, but we also want to see a significant pairwise comparison for B at the level of A known to show the effect, and no significant pairwise comparison at the level known not to have the effect. This requires extra participants. Figure 5 shows how often we obtain the expected pattern as a function of total sample size. As you can see, there is much noise below sample sizes of 100. Even above this sample size you do not always find the expected pattern, mostly because there is an effect of B at the A level where no effect is expected.

Multiple stimuli per condition

So far, the discussion was limited to designs with one (summary) variable per participant per condition. In bilingualism research we often have many observations per condition and we want to generalize across stimuli as well as across participants. For instance, if we want to compare the word frequency effect in first and second language, we will present more than one low-frequency word and one high-frequency word in each language. The analysis of such datasets is increasingly done with linear mixed effects (LME) models.

At present, there is a dearth of information on the power of designs with multiple observations per participant per condition (see Brysbaert & Stevens, 2018) and the present short report precludes further discussion. However, simulations suggest that, as a general rule of thumb, the numbers mentioned so far also work for reaction time studies with 40 or more observations per condition.

Discussion

In the introduction we argued that investigating scientific issues with underpowered studies is like looking at scenes with bad lenses (Figure 1). It increases the weight of interpretation over that of observation. As a result, statistical tests lose their power to decide between likely and unlikely hypotheses and are reduced to a rhetoric prop, shoring up claims that look sensible to the researchers (and the reviewers).

The situation looks particularly dire for between-groups comparisons and for interactions. For these effects it is to be feared that a substantial percentage of significant findings published in the literature are false alarms due to an alpha rate of 5%. The risk is augmented by the fact that complex designs easily include several interaction effects, so that false positives are prevalent if no correction for multiple testing is made (analyses involving 20 interaction terms are on average expected to yield one significant effect on the basis of sampling error alone). The risk may further be augmented by the use of questionable research practices and the fact that researchers often have considerable freedom in which dependent variables to analyze and which analyses to use (Gelman & Loken, 2013; Von der Malsburg & Angele, 2017).

Whereas the probability of false alarms is very similar for main effects and interaction effects, obtaining a genuine effect requires many more participants for interactions than for main effects. A sensible rule of thumb is four times as many. This means that genuine interaction effects will often be insignificant in studies with small numbers of participants and remain undetected if the researcher has no particular interest in them. This is particularly true for interactions with a between-groups variable.

Finding a significant interaction is one thing, being able to replicate it is another, because what we want is to replicate the SAME PATTERN OF EFFECTS. If the significant interaction was due to a significant effect at A1 and not at A2, we want to replicate not only a significant interaction, but also the same pattern of effects. This is particularly a problem for complicated, higher-order interactions. Herzog, Francis, and Clarke (2019, pp. 91-93) illustrate how the power of exact replications of complex interactions can be rather low and sometimes cannot be improved by running extra participants.

Given what we know now, it is clear that we have to step up our game if we want research on bilingualism to be more than an endless quarrel about exciting, new, significant observations that others find difficult to replicate. The solutions are not overly complicated; they just require us to organize our work differently (see also Brysbaert, 2019). These are some suggestions.

- Keep your design as simple as possible. Each extra variable multiplies the number of participants you have to test. This is particularly important if you are testing a small or difficult to reach population.
- Organize the work so that more participants can be tested, for instance by collaborating with many labs (ManyBabies Consortium, 2020) or by using online testing (Nichols, Wild, Stojanoski, Battista & Owen, 2020).
- If the data are variable (e.g., reaction times), test participants more thoroughly, so that you get reliable effects at the participant level.
- Be happier with one properly powered study than with 10 underpowered studies, which mainly increase the noise in the literature.
- Do not accept hopelessly underpowered studies as reviewer or editor, even though the finding is exciting and was predicted by the authors. Ask for a well-powered, preregistered replication, which you will publish independent of the outcome.

Supplementary materials. A file describing the simulations with R code to reproduce them is available at https://osf.io/t7f2n/.

References

Brunner J and Schimmack U (2020) Estimating Population Mean Power Under Conditions of Heterogeneity and Selection for Significance. *Meta-Psychology*, **4**, *MP*.2018.874. DOI: https://doi.org/10.15626/MP.2018.874

Brysbaert M (2019) How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, **2**, 16. DOI: http://doi.org/10.5334/joc.72

Brysbaert M and Stevens M (2018) Power Analysis and Effect Size in Mixed Effects Models: A Tutorial. *Journal of Cognition*, **1**, 9. DOI: http://doi.org/10.5334/joc.10

De Bruin A, Treccani B and Della Sala S (2015) Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science*, **26**, 99–107.

Funder DC and Ozer DJ (2019) Evaluating effect size in psychological research: Sense and nonsense. *Advances in Methods and Practices in Psychological Science*, **2**, 156–168.

Gelman A and Carlin J (2014) Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, **9**, 641–651.

Gelman A and Loken E (2013) *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Department of Statistics, Columbia University. Available at https://osf.io/n3axs.

Herzog MH, Francis GS and Clarke A (2019) *Understanding Statistics and Experimental Design: How to Not Lie with Statistics*. Springer. Available at https://link.springer.com/content/pdf/10.1007/978-3-030-03499-3.pdf

John LK, Loewenstein G and Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, **23**, 524–532.

Judd CM, McClelland GH and Ryan CS (2008) *Data analysis: A model comparison approach* (2nd edition). Routledge.

Kerr NL (1998) HARKing: Hypothesizing After the Results Are Known. *Personality and Social Psychology Review*, **2**, 196–217. http://dx.doi.org/10.1207/s15327957pspr0203_4

Kim JY (2020) Discrepancy between heritage speakers' use of suprasegmental cues in the perception and production of Spanish lexical stress. *Bilingualism: Language and Cognition*, 1–18. Advance publication available at: DOI: https://doi.org/10.1017/S1366728918001220

LeBel EP, Campbell L and Loving TJ (2017) Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*, **113**, 230–243.

Lehtonen M, Soveri A, Laine A, Järvenpää J, De Bruin A and Antfolk J (2018) Is bilingualism associated with enhanced executive functioning in adults? A meta-analytic review. *Psychological Bulletin*, **144**, 394.

Loiselle D and Ramchandra R (2015) A counterview of 'An investigation of the false discovery rate and the misinterpretation of p-values' by Colquhoun (2014). *Royal Society Open Science*, **2**, 150217. DOI: https://doi.org/10.1098/rsos.150217

ManyBabies Consortium. (2020) Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*, **3**, 24–52.

Maxwell SE, Lau MY and Howard GS (2015) Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, **70**, 487–498.

McElreath R and Smaldino PE (2015) Replication, Communication, and the Population Dynamics of Scientific Discovery. *PloS One*, **10**, e0136088. https://doi.org/10.1371/journal.pone.0136088.

Nichols ES, Wild CJ, Stojanoski B, Battista ME and Owen AM (2020) Bilingualism Affords No General Cognitive Advantages: A Population Study of Executive Function in 11,000 People. *Psychological Science*. Preprint avaialable at https://doi.org/10.1177/0956797620903113

Paap K, Mason L, Zimiga B, Silva Y and Frost M (2020) The Alchemy of Confirmation Bias Transmutes Expectations into Bilingual Advantages: A Tale of Two New Meta-Analyses. *Quarterly Journal of Experimental Psychology*. Preprint available at DOI: 10.1177/1747021819900098.

Perugini M, Gallucci M and Costantini G (2018) A Practical Primer to Power Analysis for Simple Experimental Designs. International *Review of Social Psychology*, **31**, 20. DOI: https://doi.org/10.5334/irsp.181

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

Shrout PE and Rodgers JL (2018) Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, **69**, 487–510.

Simmons JP, Nelson LD and Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, **22**, 1359–1366. DOI: https://doi.org/10.1177/0956797611417632

Simmons JP, Nelson LD and Simonsohn U (2018) False-positive citations. *Perspectives on Psychological Science*, **13**, 255–259.

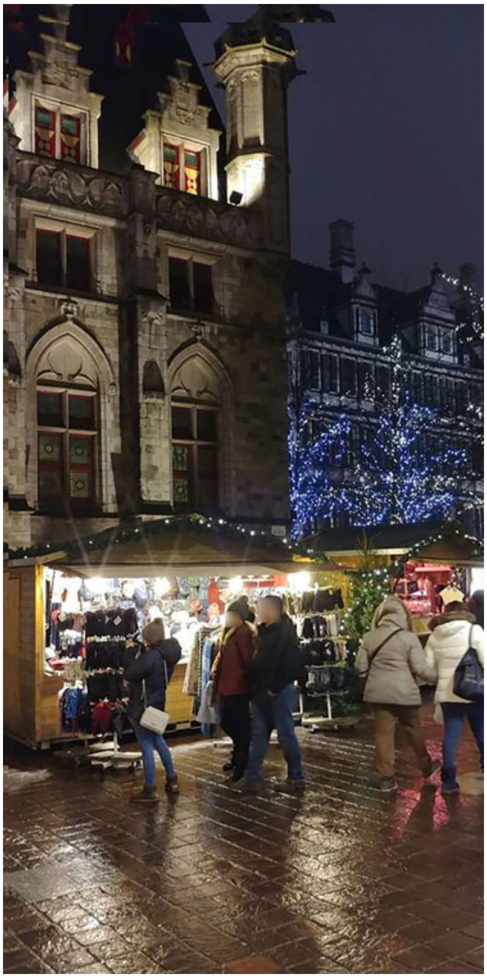
Simonsohn U (2014, March 12). *No-way interactions* [Blog post]. Retrieved from http://datacolada.org/17. DOI: https://doi.org/10.15200/winn.142559.90552

Sterne J, Becker B and Egger M (2005) The funnel plot. In Rothstein HR, Sutton AJ and Borenstein M (eds), *Publication Bias in Meta-Analysis: Prevention, Assessment and Adjustments*. John Wiley and Sons, pp. 75–98.

Vasishth S, Mertzen D, Jäger LA and Gelman A (2018) The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, **103**, 151–175.

Von der Malsburg T and Angele B (2017) False positives and other statistical errors in standard analyses of eye movements in reading. *Journal of Memory and Language*, **94**, 119–133.

Appendix:



Higher resolution image of Figure 1 (Christmas market in Ghent)