KU LEUVEN

FACULTY OF PSYCHOLOGY AND
EDUCATIONAL SCIENCES

Quantitative Psychology and Individual Differences

Towards better research practices in psychology

Sara Steegen

Doctoral thesis offered to obtain the degree of
Doctor of Psychology (PhD)

Supervisor: Prof. Dr. Wolf Vanpaemel
Co-supervisor: Prof. Dr. Francis Tuerlinckx

2017

Sara Steegen. **Towards better research practices in psychology**. Dissertation submitted to obtain the degree of PhD in Psychology, October 2017. Supervisor Prof. Dr. Wolf Vanpaemel. Co-supervisor Prof. Dr. Francis Tuerlinckx.

Psychology faces a deep crisis of confidence and is at the risk of losing its credibility. Researchers are being criticized for the way they are conducting studies, analyzing data and reporting results. Confronted with this poor research quality in psychology, several recommendations have been made to overcome this problem. The goal of this dissertation is to make a contribution to this enterprise of improving research quality in the field of psychology.

In **Chapter1**, we carry out a replication study, implementing the most commonly made recommendation for good research practices. In particular, we aim to replicate the crowd within effect, according to which the average of two guesses from one person provides a better estimate than the single guesses on their own. We also tried to make this an exemplary study, in the sense that we attempted to follow most recommended good research practices when carrying out the study.

In **Chapter2**, we extend the class of recommendations that focus on transparency by highlighting the importance of an increased transparency about arbitrary choices in data processing. We start from the observation that processing raw data into a data file ready for analysis often involves arbitrary choices among several reasonable options for excluding, transforming, and coding data. Using a worked example focusing on the effect of fertility on religiosity and political attitudes, we show that these arbitrary choices can lead to widely fluctuating results. We suggest that instead of performing only one analysis, researchers should perform a multiverse analysis, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data processing and gives pointers as to which choices are most consequential in the fragility of the result.

Chapters 3 and 4 cover topics concerning Bayes factors, which are being advocated as a Bayesian alternative for null hypothesis significance testing.

In **Chapter 3**, we compare the Bayes factor with an alternative Bayesian model selection method: the Prior Information Criterion (PIC). This latter method is a recently developed Bayesian model selection method, with close resemblances to the Bayes factor. Both methods are compared on their behavior in the context of the binomial model and we derive formal relations between them. We show that the PIC can lead to conclusions that not only widely differ from the conclusions based on the Bayes factor, but are also highly undesirable.

Finally, in **Chapter 4**, we extend the core idea of Bayes factors — considering average fit rather than best fit — to qualitative data. Whereas Bayes factors focus on fit with respect to the quantitative aspects of the data, psychologists are often interested in the qualitative aspects of the data, such as ordinal patterns. We explore the potential of Parameter Space Partitioning — a model evaluation tool that focuses on qualitative data patterns — as a model selection method, focusing on average model fit with respect to the qualitative aspects of the data.

Sara Steegen. **Op weg naar betere onderzoekspraktijken binnen de psychologie**. Proefschrift aangeboden tot het verkrijgen van de graad van Doctor in de Psychologie, October 2017. Promotor Prof. Dr. Wolf Vanpaemel. Co-promotor Prof. Dr. Francis Tuerlinckx.

Psychologie ondergaat een vertrouwenscrisis en loopt het risico om zijn geloofwaardigheid te verliezen. Onderzoekers worden bekritiseerd voor de manier waarop ze studies uitvoeren, data analyseren, en hun resultaten rapporteren. Als reactie op deze confrontatie met een slechte onderzoekskwaliteit in de psychologie zijn er verschillende aanbevelingen voorgesteld om dit probleem op te lossen. Het doel van dit proefschrift is om een bijdrage te leveren aan deze onderneming om de onderzoekskwaliteit binnen het veld van de psychologie te verbeteren.

In **Hoofdstuk 1** voeren we een replicatiestudie uit. Daarmee implementeren we de meest geciteerde aanbeveling voor goede onderzoekspraktijken. Meer specifiek willen we het "crowd within" effect repliceren, wat inhoudt dat het gemiddelde van twee gokken van een persoon een betere schatting geeft dan de eerste en de tweede gok apart. We probeerden van deze studie ook een voorbeeldstudie te maken, in de zin dat we de meest frequente aanbevelingen voor goede onderzoekspraktijken trachtten te volgen bij het uitvoeren van de studie.

In **Hoofdstuk 2** breiden we de categorie van aanbevelingen die focussen op transparantie uit door het belang van een verhoogde transparantie over arbitraire keuzes in dataverwerking te benadrukken. We vertrekken vanuit de observatie dat het verwerken van ruwe data tot een databestand dat geschikt is voor data analyse vaak arbitraire beslissingen tussen verschillende redelijke opties voor het verwijderen, transformeren en coderen van data met zich meebrengt. Aan de hand van een uitgewerkt voorbeeld dat focust op het effect van vruchtbaarheid op religiositeit en politieke attitudes, tonen we dat deze arbitraire keuzes kunnen leiden tot ruim fluctuerende resultaten. We suggereren dat in de plaats van het uitvoeren van een enkele analyse, onderzoekers een multiverse analyse zouden moeten uitvoeren, wat inhoudt dat alle analyses over de hele set van alternatief verwerkte data sets worden uitgevoerd, corresponderend aan een grote set van aannemelijke scenario's. Een multiverse analyse geeft een idee over de mate waarin conclusies veranderen door arbitraire keuzes in dataverwerking, en geeft aanwijzingen over welke keuzes het meest invloedrijk zijn op de kwetsbaarheid van het resultaat.

Hoofdstukken 3 en 4 gaan over onderwerpen die te maken hebben met Bayes factoren, wat bepleit wordt als Bayesiaans alternatief voor nulhypothese significantietoetsing.

In **Hoofdstuk 3** vergelijken we Bayes factoren met een alternatieve Bayesiaanse modelselectiemethode: de Prior Information Criterion (PIC). Deze laatste methode is een recent ontwikkelde Bayesiaanse modelselectiemethode, met grote gelijkenissen met de Bayes factor. Beide methoden worden vergeleken met betrekking tot hun gedrag in de context van het binomiaal model, en we leiden formele relaties af tussen hun. We tonen dat de PIC kan leiden tot conclusies die niet alleen ruim verschillen van die gebaseerd op de Bayes factor, maar ook zeer onwenselijk zijn.

Tot slot, in **Hoofdstuk 4** breiden we het basisidee van de Bayes factor — de gemiddelde fit in plaats van de beste fit beschouwen — uit naar kwalitatieve data. Terwijl Bayes factoren focussen op fit met betrekking tot de kwantitatieve aspecten van de data, zijn psychologen vaak geïnteresseerd in de kwalitatieve aspecten van de data, zoals ordinale patronen. We onderzoeken het potentieel van Paramater Space Partitioning — een modelevaluatie-instrument dat focust op kwalitatieve datapatronen — als een modelselectiemethode, met een focus op de gemiddelde model fit met betrekking tot de kwalitatieve aspecten van de data.

Dankwoord

Na zeven jaar neem ik afscheid van het PSI. De tristesse die het gebouw uitstraalt staat in schril contrast met de leuke herinneringen die ik er heb opgebouwd.

Wolf, je was niet enkel mijn promotor, maar ook mijn OKP-peter. De 'scenic tour' die vasthangt aan die laatste functie heb ik nooit van jou gekregen, maar dit heb je ruimschoots goedgemaakt doordat ik met al mijn vragen bij jou terechtkon. Ik voelde me altijd door jou gesteund. Je hebt mijn onderzoek in goede banen geleid, mij steeds geholpen als ik ergens mee vastzat, nauwkeurig de onderzoekskwaliteit bewaakt en je bleef me tot op het einde herinneren aan praktische zaken die ik moest regelen en vaak vergat. We waren niet het meest georganiseerde duo, maar voor mij was het een heel leuke samenwerking die ik zeker ga missen. Bedankt om zo'n goede promotor te zijn.

Francis, bedankt om zo'n goede co-promotor te zijn. Je was heel betrokken bij mijn onderzoek, steeds van alles op de hoogte. De vergaderingen met jou erbij waren gezellig. Soms wat te gezellig, want er werd wel eens afgedwaald. Je stond altijd klaar met interessante ideeën, nieuwe inzichten, en een stevige portie relativeringsvermogen. Je flauwe moppen nam ik er graag bij.

Dries, hoeveel ik je ook in de toekomst nog zal tegekomen, ik zal je nooit los zien van bureau C01.64. Ik zal je nooit los zien van alle planten die we hebben zien sterven, van de leuke plaatjes die we luisterden, van de verwarming die ik van jou nooit te hoog mocht zetten, van de nootjes die je met iedereen deelde, en van de talloze attributen die niet thuishoorden in een werkomgeving maar er toch hun vaste plek hadden gevonden (strijkplank, bak bier, slaapzak, skateboard, ...). Ik heb je gemist het laatste jaar.

Annelies, ook jij bent niet weg te denken uit mijn C01.64 herinneringen. De koekjestrommel, de bordtekeningen en de extra stem om de verwarming hoger te zetten, waren van grote meerwaarde voor mij.

Aniek, bedankt om me in het laatste jaar zo warm te onhalen in bureau B01.89. Ik stond dan wel onderaan de ladder ('de slechte plek gaat naar de nieuwkomers, de goede plek naar de anciens'), maar ik heb me er altijd thuisgevoeld, dankzij jou.

Alle ander collega's van OKP/QuPID, bedankt voor de leuke jaren. Ik heb collega's zien komen, ik heb collega's zien gaan, maar de gemoedelijke OKP-sfeer bleef steeds bestaan. Bedankt aan alle collega's van het didactisch team van Statistiek II, voor het delen van ervaringen over lesgeven, een schorre stem en warme practicumlokalen. Jasmine, bedankt voor de koffie en de vrolijke begroetingen elke ochtend. Eva, Stijn, Tom, Kristof, Isabelle, Maddy, Merijn, bedankt voor de fijne feestjes en muziekquizzen. Laura en Lisa, bedankt voor de gezelligheid, ook jullie heb ik gemist het afgelopen jaar.

Vrienden en familie, bedankt om telkens genoegen te nemen met een kort antwoord op jullie vragen over mijn doctoraat, aangezien ik meestal geen zin had om hier uitgebreid over te vertellen.

Mama, papa en Ruben, bedankt voor alle hulp en steun de afgelopen jaren.

Andreas, jij hebt mijn OKP-jaren op verschillende manieren opgefleurd. Ik denk in de eerste plaats aan de prachtige kerstboom die je binnenbracht in onze bureau; een versiertruc die zijn werk heeft gedaan. Sindsdien kwam je te pas en te onpas binnenwaaien voor een babbel. Meestal vond ik dat leuk en als dat niet zo was, dan wendde je je gelukkig tot mijn collega's. Met de komst van Polly heb je me ook van de ultieme stressafleider voorzien. Bedankt!

Contents

General Introduction

Crisis of confidence

Psychology faces a deep crisis of confidence and is at the risk of losing its credibility. Researchers are being criticized for the way they are conducting their studies, analyzing their data and reporting their results. When collecting and analyzing their data, researchers have the tempting opportunity to use strategic practices that lead to publishable results but jeopardize the trustworthiness of the findings. One such procedure is sneaking in exploratory strategies in a confirmatory testing procedure, in which researchers explore various alternatives and selectively report only the statistically significant results with post-hoc explanations, portraying them as confirmatory findings (e.g. Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011). This can lead to biased effect size estimates and increased rates of false positive findings. As a result, many findings in psychological science are hard to replicate (e.g., Open Science Collaboration, 2015). Although questionable research practices have been acknowledged and have been warned for years ago already (De Groot, 1956/2014; Kerr, 1998), it is only recently that psychology is in a heated discussion about this issue (e.g., Pashler & Wagenmakers, 2012; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkielman, & Pashler, 2009). It seems to be not uncommon that psychologists engage in dubious research practices, such as failing to report all dependent measures, neglecting negative outcomes, collecting more data af-

ter seeing whether the results were significant, and so on (John, Loewenstein, & Prelec, 2012).

Improving research practices

Confronted with this poor research quality in psychology, several recommendations have been made to overcome this problem. One of the most cited recommendations is replication: The most straightforward way for gaining confidence in findings is by confirming them in replication studies. Some authors have emphasized the need for direct replication studies (e.g., Brandt et al., 2014; LeBel & Peters, 2011; Simons, 2014), in which the methods and procedures of the original study are copied as close as possible, whereas others have argued for conceptual replications (e.g., Stroebe & Strack, 2014), in which original study aspects are deliberately changed in order to evaluate the generalizability of the effect across different conditions.

A major class of recommendations involves an increased transparency. For example, by clearly reporting which findings are confirmatory and exploratory, readers get a more fair view on the results. A procedure that automatically imposes this division is preregistration, which involves registering the study hypotheses, methods and analyses before data collection. One way in which researchers can preregister their study is via online platforms such as the Open Science Framework (Spies et al., 2012), which offers researchers a user-friendly environment for specifying their study details and "freezing" or fixing this information before data collection. Another form of pre-registration is through a "registered report", which is a publishing format in which a researcher's hypotheses and methodology are submitted and reviewed before data collection. Once this registered report is accepted, and the data are collected and analyzed according to this report, the study is published, independent of the results. Many researchers have been advocating preregistration (e.g., Asendorpf et al., 2013; Ioannidis, 2014; Wagen-

makers, Wetzels, Borsboom, van der Maas, & Kievit, 2012) and more and more journals support the possibility for submitting registered reports (e.g., Chambers, 2013; Nosek & Lakens, 2016). Other much discussed topics in the class of transparency recommendations include open data and materials (e.g., Nosek et al., 2015; Sijtsma, 2016; Rouder, 2016; Wicherts, Bakker, & Molenaar, 2011), and the disclosure of information about sample size determination, data exclusion, all manipulations and all measures (Simmons, Nelson, & Simonsohn, 2012).

Another class of recommendations is concerned with a change in analysis methods. Most psychologists rely on p-values to draw conclusions about their hypotheses: Results are significant, whenever a pre-specified significance level is reached. Some researchers argue against this dichotomous way of inference and suggest to rely on estimation, such as effect sizes and confidence intervals: Instead of focusing on the *presence* of an effect, it is considered to be more informative to focus on the *size* of the effect (e.g., Asendorpf et al., 2013; Cumming, 2013). In fact, this guideline has already been proposed almost two decades ago by the American Psychological Association Task Force on Statistical Inference (Wilkinson, 1999), and it has now revived in the current crisis. Other researchers are advocating to move away from the traditional frequentist statistical framework, and adopt the Bayesian framework instead, with Bayes factors and posterior distributions as Bayesian equivalents for hypothesis testing and estimation, respectively (e.g., Dienes, 2016; Kruschke, 2011; Rouder, Speckman, Sun, Morey, & Iverson, 2009; Wagenmakers et al., 2011).

Other recommendations are, for example, situated at the level of study design (e.g., perform a power analysis, use more strong manipulations, Simmons et al., 2011; Schimmack, 2012), data collection (e.g., blinding of experimenters, Wicherts et al., 2016), analysis protocols (e.g., blind analysis, using a co-pilot approach, MacCoun & Perlmutter, 2015; Wicherts, 2011),

the review process (e.g., peer reviewer's openness initiative, transparency and openness promotion guidelines, Morey et al., 2016; Nosek et al., 2015), academic incentive structures (e.g., more emphasis on quality, rather than quantity, of publications, Ioannidis, Munafò, Fusar-Poli, Nosek, & David, 2014) and education (e.g., establish a standard of good research practice in the classroom, Asendorpf et al., 2013).

Clearly, in the light of the crisis of confidence, psychology has been subject to a wide range of recommendations, all sharing the same goal: Improving research quality in order to restore confidence. The goal of this dissertation is to make a contribution to this movement towards better research practices in psychology.

Overview of the chapters in this dissertation

The dissertation starts with a replication study (**Chapter 1**), directly addressing the most commonly made recommendation for good research practices. In particular, we aimed to replicate the *crowd within effect*, according to which the average of two guesses from one person provides a better estimate than the single guesses on their own (Vul & Pashler, 2008). In this study, we attempted to follow good research practices when carrying out the study. For example, before data collection, we preregistered our study, including the hypotheses, materials and analysis code. Concerning data analysis, we evaluated our findings from different perspectives (i.e., null hypothesis significance testing, effect sizes, confidence intervals, and Bayes factor). We adopted a co-pilot, multi-software approach, in the sense that all analyses were performed independently by two researchers using different analysis software. Finally, we made all our study information publicly available on the Open Science Framework.

In **Chapter 2**, we extend the class of transparency recommendations by highlighting the importance of an increased transparency about arbitrary

choices in data processing. We start from the observation that processing raw data into a data file ready for analysis often involves arbitrary choices among several reasonable options for excluding, transforming, and coding data. Using a worked example focusing on the effect of fertility on religiosity and political attitudes, we show that these arbitrary choices can lead to widely fluctuating results. We suggest that instead of performing only one analysis, researchers should perform a *multiverse analysis*, which involves performing all analyses across the whole set of alternatively processed data sets corresponding to a large set of reasonable scenarios. A multiverse analysis offers an idea of how much the conclusions change because of arbitrary choices in data processing and gives pointers as to which choices are most consequential in the fragility of the result.

In the remaining two chapters, we focus on one particular recommendation for better research practices in specific: The adoption of the Bayesian, rather than the frequentist, statistical framework. Chapters 3 and 4 cover topics concerning Bayes factors, which are being advocated as a Bayesian alternative for p-values.

In **Chapter 3**, we compare the Bayes factor with an alternative Bayesian model selection method: the Prior Information Criterion (PIC; van de Schoot, Hoijtink, Romeijn, & Brugman, 2012). This latter method is a recently developed Bayesian model selection method, with close resemblances to the Bayes factor. Both methods are compared on their behavior in the context of the binomial model and we derive formal relations between them. We show that the PIC can lead to conclusions that not only widely differ from the conclusions based on the Bayes factor, but are also highly undesirable.

Finally, in **Chapter 4**, we extend the core idea of Bayes factors — considering average fit rather than best fit — to qualitative data. Whereas Bayes factors focus on fit with respect to the quantitative aspects of the data, psychologists are often interested in the qualitative aspects of the data, such

as ordinal patterns. We explore the potential of Parameter Space Partitioning (PSP; Pitt, Kim, Navarro, & Myung, 2006) — a model evaluation tool that focuses on qualitative data patterns — as a model selection method, focusing on average model fit with respect to the qualitative aspects of the data.

Each of the four chapters corresponds to a published paper:

Chapter 1: Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology, 5*, 786.

Chapter 2: Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*, 702-712.

Chapter 3: Steegen, S., Woojae, K., Pestman, W., Tuerlinckx, F, & Vanpaemel, W. (in press). A theoretical note on the prior information criterion. *Journal of Mathematical Psychology.*

Chapter 4: Steegen, S., Tuerlinckx, F., & Vanpaemel, W. (2017). Using parameter space partitioning to evaluate a model's qualitative fit. *Psychonomic Bulletin Review, 24*, 617-631.

Chapter 1

Measuring the crowd within again: A pre-registered replication study

Introduction

Quantitative judgements to general knowledge questions are famously known to be more accurate when estimates are averaged over a crowd compared to the individual estimates (Surowiecki, 2004). When individuals are guessing independently of each other, the crowd's estimate will be closer to the truth than the majority of the individual guesses. This "wisdom of the crowds" effect has been observed in a wide range of applications, including weight guessing, ordering tasks and market predictions (e.g., Dani, Madani, Pennock, Sanghai, & Galebach, 2012; Galton, 1907; Steyvers, Lee, Miller, & Hemmer, 2009).

In an elegant experiment, Vul and Pashler (2008) showed that the wisdom of crowds can also be obtained within a single individual. Participants were asked to make their best guess on eight general knowledge questions. Immediately after completing this, participants were unexpectedly asked to make a second, different guess for each question. The results showed that, overall, the average of two estimates from one person was more accurate than a single estimate of that person. It thus seems as if people possess a

crowd within they can consult.

Vul and Pashler (2008) further showed that increasing the independence between both guesses strengthens the crowd within effect. In particular, a second group of participants was asked to answer the same questions three weeks later instead of immediately after completing the first questionnaire. The benefit of averaging two guesses within a person was larger in this *delayed* condition than when the second guess was elicited immediately (i.e., the *immediate* condition).

On a practical level, the phenomenon that averaging multiple guesses within a person improves estimation accuracy has useful implications in daily life with respect to decision making, as it shows that judgements can benefit from the proverbial "sleeping on it". The crowd within effect has also important theoretical implications, as it suggests that our knowledge is represented in internal probability distributions from which responses are sampled. As such, it provides a solid ground of evidence for the emerging idea that human reasoning rests on Bayesian inference (e.g., Jones & Love, 2011; Tenenbaum, Griffiths, & Kemp, 2006). The practical and theoretical appeal of the crowd within has resulted in extensive media coverage (e.g., Herbert, 2008; "The Crowd Within", 2008) and 78 citations (according to Google Scholar, June 7, 2013).

We know of four studies that attempted to replicate the crowd within effect in the immediate condition, with mixed results. Two of these studies report the finding that averaging two successive guesses from one person provides better estimates than the single guesses (Hourihan & Benjamin, 2010; Rauhut & Lorenz, 2011)[1]. The results from the two remaining studies were somewhat mixed. In line with the crowd within effect, Herzog and

[1]In Hourihan and Benjamin (2010), a minority of the subjects made the second guess after a delay of less than an hour instead of immediately following the first guess. However, Hourihan and Benjamin (2010) reported that this procedural difference did not affect the data, so it was not examined further.

Hertwig (2009) showed that aggregating two guesses improved estimation accuracy compared to single guesses. However, the 95% confidence interval for this accuracy gain included the null value[2]. Finally, Edward Vul informed us about an unpublished replication attempt that failed to find a significant improvement of the average of two guesses compared to the first guess (Banker & McCoy, 2013). However, in support of the crowd within, the results did point in the expected direction (see Table 1.1 for more details).

To the best of our knowledge, there are no replication attempts of the delayed condition. This condition yielded the strongest effect of the crowd within, which is in line with the idea that the benefit of averaging is a result of the different guesses being sampled from a probability distribution. Since a certain level of independence between the errors of the estimates is crucial in order to get this effect, it is logical that a three-week delay between the guesses (inducing a greater independence) enhances the benefit. Although this manipulation has never been adopted in other research studying the crowd within, a few studies did show that other factors boosting independence between the two guesses enhance the benefit of averaging (e.g., Herzog & Hertwig, 2009; Hourihan & Benjamin, 2010).

In the light of the practical and theoretical appeal of the crowd within, and the limited success in replicating the effect, we believe it is worthwhile to set up another attempt at replicating the crowd within effect in both the immediate and the delayed condition of Vul and Pashler (2008).

[2]More specifically, the reported accuracy gain (defined as the median decrease in error of the average of the two estimates relative to the first estimate, across items) was on average 0.3 percentage points with SD = 2.3%, Mdn = 0%, CI = [0.0% - 0.8 %] and d = 0.12.

Method

Before the start of the data collection, we preregistered this study at the Open Science Framework (`osf.io/p2qfv;` Spies et al., 2012). We preregistered a document containing details about samples size calculation, recruitment plan, materials and procedure, data cleaning, and data analysis. We also preregistered details of the experiment such as the instructions, and an executable Matlab script of the confirmatory analyses. This code can also be found in Appendix A. This whole method section (until the results section) is literally copied from the preregistered document, except for changed tenses, updated references, and added footnotes 4 and 5.

Sampling Plan

Immediate condition

The sampling plan for the immediate condition was based on a power analysis considering the existing evidence for the crowd within effect in this condition from the original paper by Vul and Pashler (2008) on the one hand, and from two replication attempts on the other hand, namely the study by Hourihan and Benjamin (2010) and the study by Banker and McCoy (2013). The results from Herzog and Hertwig (2009) were not included in the power analysis, because these authors measured accuracy gain in a different way than Vul and Pashler (2008)[3], making the reported statistics in these two studies incomparable. Finally, the results from Rauhut and Lorenz (2011) could not be considered, since this study did not report sufficient information to calculate the required effect sizes.

[3]Whereas Vul and Pashler (2008) measured accuracy gain by taking the difference between the mean squared error (MSE) of one of the single estimates and the MSE of the aggregated guess, Herzog and Hertwig (2009) defined accuracy gain as the median decrease in absolute error of the aggregated estimates relative to the first estimate.

The power calculations were based on a weighted average effect size across the three relevant studies. As a measure of effect size, we used Cohen's standardized mean difference d_z for dependent groups (Cohen, 1988, p. 48):

$$d_z = \frac{\mu_X - \mu_Y}{\sigma_{X-Y}} = \frac{\mu_X - \mu_Y}{\sqrt{\sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y\rho_{XY}}}, \tag{1.1}$$

where μ_X and μ_Y are the means in the two groups, σ_X and σ_Y are the standard deviations in the two groups and ρ_{XY} is the correlation between the pairs of observations. This effect size can easily be estimated from the t-statistic for dependent groups for a given effect and the corresponding sample size n, as follows:

$$\hat{d_z} = \frac{t}{\sqrt{n}}. \tag{1.2}$$

Table 1.1 shows the means, standard deviations, correlations between the pairs of observations, sample sizes, t-statistics and p-values of the immediate condition in all three studies, as well as the resulting effect sizes. Since the crowd within effect comprises an accuracy gain of the average guess compared to either of both single guesses, two effect sizes were calculated in each study: One effect size for guess 1 (i.e., the standardized mean difference between the mean squared error (MSE) of guess 1 and the MSE of the aggregated guess) and another effect size for guess 2 (i.e., the standardized mean difference between the MSE of guess 2 and the MSE of the aggregated guess).

We pooled the individual effect sizes across the studies by weighing each effect size with its inverse variance (Cooper, Hedges, & Valentine, 2009):

$$\overline{d_z} = \frac{\sum_{i=1}^{k} w_i \hat{d}_{z_i}}{\sum_{i=1}^{k} w_i}, \tag{1.3}$$

where k is the number of studies and w_i is the inverse of the variance v_i of effect size $\hat{d}_{z_i}$ in study i:

$$\frac{1}{w_i} = v_i = \left(\frac{1}{n_i} + \frac{\hat{d}_{z_i}^2}{2n_i}\right) 2(1 - r_i), \tag{1.4}$$

where n_i is the sample size and r_i is the correlation between the pairs of observations in study i.

Pooling the effect sizes across the three studies resulted in a weighted average effect size of $\overline{d_z} = .17$ for guess 1 and $\overline{d_z} = .56$ for guess 2.

Using G*Power 3.1 (Faul, Erdfelder, Buchner, & Lang, 2009), we calculated a planned sample size for achieving a .95 power level using a two-tailed dependent t-test, given both effect sizes. This resulted in a sample size $n = 439$ for guess 1 and $n = 31$ for guess 2.[4] To be most conservative, we planned to adopt at least the largest sample size of these two, that is $n = 439$.

Delayed condition

As there are no known replication attempts of the crowd within effect in the delayed condition, the sampling plan for this condition was based on an effect size estimated from Vul and Pashler (2008) only. Again, we used Cohen's d_z as a measure of effect size, and we estimated an effect size for guess 1 and guess 2, using formula (1.2). The estimated effect sizes, as well as the means, standard deviations, correlations, sample sizes, t-statistics and p-values are shown in Table 1.2. Using G*Power 3.1, we calculated a planned sample size $n = 48$ for guess 1 and $n = 13$ for guess 2 in order to achieve a power of .95, using a two-tailed dependent t-test.[5] To be conservative, we planned to adopt a sample size of at least $n = 48$ in the delayed condition.

[4]These target sample sizes are incorrect, and should be $n = 452$ for guess 1 and $n = 44$ for guess 2 instead. As the effective samples size, as reported in the Sample section ($n = 471$), exceeds the corrected target sample sizes, this error is non-consequential (see also Steegen, Dewitte, Tuerlinckx, & Vanpaemel, 2014).

[5]These target sample sizes are incorrect, and should be $n = 61$ for guess 1 and $n = 26$ for guess 2 instead. As the effective samples size, as reported in the Sample section ($n = 140$), exceeds the corrected target sample sizes, this error is non-consequential (see also Steegen et al., 2014).

Table 1.1: *Statistics for guess 1 and guess 2 in the immediate condition for the three studies included in the power analysis.*

Study	Mean MSE single guess	Mean MSE average guess	SD MSE single guess	SD MSE average guess	r	n	t	p	$\hat{d}_z$
				Guess 1					
Vul and Pashler (2008)	555*	508*	361*	305*	.88*	255$^{\triangle}$	4.41*	$< .001$*	.28
Hourihan and Benjamin (2010)	502$^{\triangle}$	484$^{\triangle}$	261°	268°	.91°	170$^{\triangle}$	2.15$^{\triangle}$	.03$^{\triangle}$	.16
Banker and McCoy (2013)	463*	452*	281*	268*	.95*	201$^{\triangle}$	1.71*	.09*	.12
				Guess 2					
Vul and Pashler (2008)	638*	508*	382*	305*	.83*	255$^{\triangle}$	9.90*	$< .001$*	.62
Hourihan and Benjamin (2010)	565$^{\triangle}$	484$^{\triangle}$	274°	268°	.87°	170$^{\triangle}$	7.89$^{\triangle}$	$< .001^{\triangle}$	.61
Banker and McCoy (2013)	509*	452*	298*	268*	.92*	201$^{\triangle}$	7.03*	$< .001$*	.50

Note: The reported t-values of Vul and Pashler (2008) differ from what is reported in the original paper. Contacting the authors concerning the experiment led them to uncover that the data show a stronger evidence for the crowd within effect than what had been reported in the article originally. The t-statistic values for the crowd within effect on both the first and the second guess were reported smaller than they actually are. The reported t-values here are the correct ones.

MSE = mean squared error, SD = standard deviation, r = correlation between single guess and average guess, n = sample size, t = dependent t-statistic value, p = p value of corresponding statistic, $\hat{d}_z$ = effect size.

*Computed from raw data, $^{\triangle}$numerically reported in paper, °derived from numerically reported statistics in paper.

Table 1.2: *Statistics for guess 1 and guess 2 in the delayed condition for the study included in the power analysis.*

Study	Mean MSE single guess	Mean MSE average guess	SD MSE single guess	SD MSE average guess	r	n	t	p	$\hat{d_z}$
				Guess 1					
Vul and Pashler (2008)	542*	447*	363*	273*	.84*	173$^{\triangle}$	6.22*	$< .001$*	.47
				Guess 2					
Vul and Pashler (2008)	610*	447*	380*	273*	.83*	173$^{\triangle}$	9.85*	$< .001$*	.75

Note: The reported t-values of Vul and Pashler (2008) differ from what is reported in the original paper. See Table 1.1 for further information.

MSE = mean squared error, SD = standard deviation, r = correlation between single guess and average guess, n = sample size, t = dependent t-statistic value, p = p value of corresponding statistic, $\hat{d_z}$ = effect size.

*Computed from raw data, $^{\triangle}$numerically reported in paper, $^{\circ}$derived from numerically reported statistics in paper.

Recruitment

We recruited participants at the University of Leuven. Psychology students were asked to participate in the experiment either in turn for course credits (immediate condition) or for a chance to win cinema tickets (delayed condition). Given the characteristics of psychology students, we expected a majority of female participants between the age of 18 and 23. Following the original paper, participants did not have to meet any inclusion criteria. As will become clear below, we did not know beforehand the exact number of participants, so the sample sizes computed above are minimal sample sizes. Data were only analyzed once to avoid multiple comparison issues.

For the immediate condition, we recruited participants until we had reached at least the planned sample size of 439. In particular, we made use of sessions where participants were assigned in batches. The size of each batch was largely beyond our control, making the exact sample size unknown beforehand. We used the minimum number of batches to reach the planned minimal sample size of 439. The actual sample size was expected to be larger than 439.

For the delayed condition, we made use of a pool of about 300 students attending a course. All of them were invited to participate in an experiment consisting of two sessions. Three weeks after the first session, students who participated in the first session were invited to participate in the second session. Data were used from students who participated in both sessions (i.e., data from students who only participated in the first session were discarded). Again, the actual sample size was unknown beforehand, and was expected to be larger than 48.[6]

[6]We planned that in the unlikely case that the planned sample size was not reached (i.e., there were less than 48 students participating twice), we would repeat the procedure in another course and combine the data from both courses in our analyses. However, this turned out to be unnecessary.

Materials and Procedure

Materials

The original material in the study of Vul and Pashler consisted of eight real-world knowledge questions[7], shown in Table 1.3. We adopted these questions with updated answers (derived from *The World Factbook*, Central Intelligence Agency, 2013), as shown in Table 1.3. These questions were translated in Dutch.

Procedure Immediate Condition

Participants were seated in front of computer screens in groups of 10-15 persons. At the beginning of the study, they were asked to activate full screen mode on their computer and to stay in this mode during the entire experiment. This was to prevent them from looking up the answers to the questions. Next, after signing an informed consent and providing their demographic details (sex, age and nationality), participants were given the eight general knowledge questions sequentially, with the instruction to guess the correct answers and to not look them up. In line with the original study (Edward Vul, personal communication, June 6, 2013), the eight questions were presented in randomized order and participants were prevented from going back to earlier questions. Immediately after completing the first questionnaire, participants were unexpectedly asked to make a second, different guess for each question. Again, the questions were presented in randomized order. After completion, participants were asked to indicate whether they had looked up the answers to the questions or not.

[7] We retrieved these questions from Edward Vul's webpage (Vul, n.d.).

Procedure Delayed Condition

Students attending a course were asked to participate in a short experiment later that day, on the internet, without explaining the task they would have to perform. At the time of the experiment, participants were asked to activate full screen mode on their computer and to stay in this mode during the entire experiment. They were also asked not to seek help from anyone while performing the task. In line with the original study (Edward Vul, personal communication, October 19, 2013), participants were informed that there would be a second session of the experiment three weeks later, but without giving advance notice that they would be answering the questions a second time. After signing an informed consent and providing their demographic details (sex, age and nationality), participants were given the eight general knowledge questions in the same way as in the immediate condition. After completing the questionnaire, participants were asked not to discuss the task with their companion students or other people, nor to look up the answers to the questions. Three weeks later, participants who participated in the first session of the experiment were invited per mail to participate in the second session, also on the internet. In this session, they were asked to give a second, different guess to each of the eight questions. At the end of the second session, participants were asked to indicate whether they had looked up the answers to the questions (either during the first session, during the second session or in the period between the two sessions) or not.

Known Differences from Original Study

The study differed from the original study in two aspects. Firstly, as we explained above, the real-world knowledge questions used in the study of Vul and Pashler (2008) were translated to Dutch and the answers to these questions were updated.

Secondly, we expected that our subject pool of undergraduate students

Table 1.3: *Questions used in Vul & Pashler (2008) with original answers and updated answers*

No.	Question	Answers in Vul & Pashler (2008)	Answers in current study
1	The area of the USA is what percent of the area of the Pacific Ocean?	6.3	6.3
2	What percent of the world's population lives in either China, India, or the European Union?	44.4	43.3
3	What percent of the world's airports are in the United States?	30.3	32.3
4	What percent of the world's roads are in India?	10.5	13.4
5	What percent of the world's countries have a higher fertility rate than the United states?	58	53.6
6	What percent of the world's telephone lines are in China, USA, or the European Union?	72.4	54.8
7	Saudi Arabia consumes what percentage of the oil it produces?	18.9	26.4
8	What percentage of the world's countries have a higher life expectancy than the United States?	20.3	22.4

would be less diverse than the internet-based subject pool in the original study with respect to variables such as age, ethnicity and educational level. However, we do not believe this is critical for a fair replication, since there is no a priori or theoretical reason why the crowd within effect would rely on these type of variables.

Confirmatory Analysis Plan

Data cleaning plan

Vul and Pashler (2008) did not adopt any data filtering procedures in their study (Edward Vul, personal communication, April 30, 2013). However, following the original authors' advice, we planned to exclude data from those participants which defocused the browser window running the study, as the latter may be an indication of participants looking up the answers to the questions. On the same line of reasoning, we planned to exclude data from participants who indicated that they had looked up the answers to the questions at the end of the experiment. Further, we planned to exclude data when impossible answers (i.e., percentages below zero or above hundred) or blank answers were given. In this case, both guesses for the concerning question were planned to be excluded from the analyses.

Analysis process

A complete replication of the Vul and Pashler (2008) paper includes a higher accuracy of the aggregated guess compared to the individual guesses, in both the immediate and the delayed condition.

In accordance with the original study, we assessed for each participant the accuracy of a guess with the MSE of the estimate across all eight questions. In each condition, the MSE of guess 1, guess 2, and the average of both guesses was calculated for each participant. The MSE of the average was calculated by first averaging guess 1 and guess 2, and then computing the MSE. Next, we compared the MSE of guess 1 and the MSE of the average of both guesses across participants by performing a two tailed t-test for paired observations. Further, we repeated this for the comparison of the MSE of guess 2 and the MSE of the average of both guesses. For each condition, if the observed values of both t-statistics were positive (i.e., both the MSE of guess 1 and the MSE of guess 2 were on average larger than the MSE of the aggregated guess) and a p-value smaller than .05 was obtained for both the tests, we evaluated the replication of the crowd within effect as being successful in the concerning condition.

Besides these traditional metrics for evaluating the success of a replication attempt, we calculated the effect sizes, $\hat{d}_z$, for guess 1 (i.e., the standardized mean difference between the MSE of guess 1 and the MSE of the aggregated guess) and for guess 2 (i.e., the standardized mean difference between the MSE of guess 2 and the MSE of the aggregated guess), together with their 95% confidence intervals. This allowed us to consider subtleties in the replication outcomes beyond the traditional dichotomy of failure or success of the replication attempt (see Simonsohn, 2015).

Following Vul and Pashler (2008), we performed two additional tests. First, we also compared the difference in accuracy between guess 1 and guess 2, again by performing a two tailed t-test for paired observations.

Vul and Pashler (2008) found that second guesses were less accurate than first guesses, indicating that the accuracy gain of averaging could not be attributed to subjects looking up the answers between guesses. Second, we compared the benefit of averaging in the immediate condition and the delayed condition by performing an unpaired t-test on the mean difference in error between the first guess and the average guess in the immediate condition versus the delayed condition. As discussed in the Introduction, Vul and Pashler (2008) found that the benefit of averaging was greater in the delayed condition than in the immediate condition.

Results

The raw data and processed data are available at the Open Science Framework (`osf.io/ivfu6`; Spies et al., 2012), accompanied with three Matlab scripts to execute the processing of the raw data, the pre-registered confirmatory analyses (see also Appendix A and B)[8] and additional post-hoc analyses. We adopted a co-pilot approach (Wicherts, 2011) in the sense that aside from the analyses based on this Matlab code, the second author independently post-processed and analyzed the data in SPSS, except for the calculation of the confidence intervals for the effect sizes and the calculation of the Bayes factors. The results obtained from these SPSS analyses were identical to the results from the Matlab analyses.

[8] Whereas Appendix A contains the preregistered Matlab code for the confirmatory analyses, we used the code in Appendix B. The code in Appendix B extends the code in Appendix A in that it contains code for the calculation of descriptive statistics, some effect sizes, some confidence intervals for effect sizes, and code to make the scatter histogram plots in Figure 1.2.

Sample

Immediate Condition

A total of 484 psychology students participated in the immediate condition. However, 11 of these participants did not complete the experiment, so the data from these participants were excluded in the data analysis. Following our preregistered data cleaning plan, we also excluded the data from two participants who indicated that they had looked up the answers to the questions. Further, we were planning to exclude data from those participants who defocused the browser window while running the study. Yet, due to a technical problem, the digital assessment of whether participants had defocused the browser window in the immediate condition was not reliable. Fortunately, this is not problematic in this condition, as at the time of the data collection an experimenter was present in the back of the room, ascertaining that participants did not defocus the browser window. Finally, we were planning to exclude data when impossible answers (i.e., percentages below zero or above hundred) or blank answers were given. However, as it was made impossible to provide these type of answers in the experiment, this part of the data cleaning plan did not need to be executed. Our final sample of 471 psychology students consisted of 397 women and 74 men, with a mean age of 19.2 (SD = 2.8). Note that the gender imbalance in our sample is according to our expectations.

Delayed Condition

A total of 231 psychology students participated in the first session of the delayed condition and 171 of these students also participated in the second session. We excluded the data from 9 participants who did not complete one or both sessions, the data from 21 participants who defocused the browser window while running the study and the data from one participant who

indicated that she had looked up the answers to the questions. Similar to the immediate condition, it was made impossible for participants to provide impossible or blank answers, so we did not need to exclude data based on these criteria. Our final sample of 140 participants consisted of 125 women and 15 men, with a mean age of 22.0 (SD = 3.1). Again, this gender imbalance is according to our expectations.

Confirmatory Analysis

As shown in Figure 1.1, both in the immediate and in the delayed condition, the accuracy of the aggregated guess was higher compared to the accuracy of the individual guesses (see also Table 1.4). In the immediate condition, the mean MSE of the average of the two guesses ($M = 541, SD = 313$) was smaller than both the mean MSE of guess 1 ($M = 589, SD = 336$), $t(470) = 8.69, p < .001, \hat{d}_z = .40$, 95% CI = [.31, .49] and the mean MSE of guess 2 ($M = 615, SD = 351$), $t(470) = 10.26, p < .001, \hat{d}_z = .47$, 95% CI = [.38, .57]. Likewise, in the delayed condition, the mean MSE of the average of the two guesses ($M = 467, SD = 260$) was smaller than both the mean MSE of guess 1 ($M = 517, SD = 288$), $t(139) = 4.02, p < .001, \hat{d}_z = .34$, 95% CI = [.17, .51] and the mean MSE of guess 2 ($M = 589, SD = 327$), $t(139) = 8.48, p < .001, \hat{d}_z = .72$, 95% CI = [.53, .90]. Thus, our results are comparable to the results obtained by Vul and Pashler (2008).

According to the traditional standards for evaluating replication attempts, the current study can be considered as a successful replication of the crowd within effect, in both the immediate and the delayed condition. Another strategy for evaluating replication attempts has recently been proposed by Simonsohn (2015), who suggests to compare confidence intervals for effect sizes with the small effect size $d_{33\%}$, associated with a power of 33% in the original study. According to this *detectability* approach, a replication attempt is successful when the null hypothesis is rejected and the effect size

Table 1.4: *Statistics for guess 1 and guess 2 in the immediate condition and the delayed condition in the current study.*

Condition	Mean MSE single guess	Mean MSE average guess	SD MSE single guess	SD MSE average guess	r	n	t	p	$\hat{d}_z$
			Guess 1						
Immediate condition	589	541	336	313	.93	471	8.69	$< .001$	.40
Delayed condition	517	467	288	260	.86	140	4.02	$< .001$	.34
			Guess 2						
Immediate condition	615	541	351	313	.90	471	10.26	$< .001$	.47
Delayed condition	589	467	327	260	.86	140	8.48	$< .001$	.72

Note: MSE = mean squared error, SD = standard deviation, r = correlation between single guess and average guess, n = sample size, t = dependent t-statistic value, p = p value of corresponding statistic, $\hat{d}_z$ = effect size.

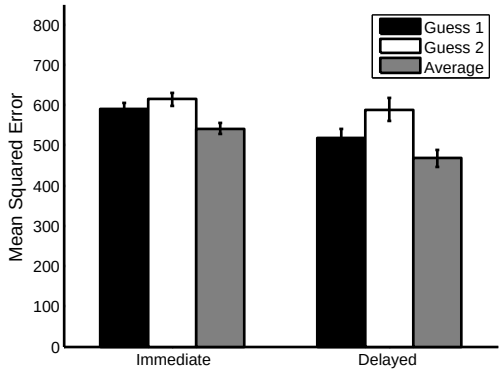


Figure 1.1: *Mean mean squared errors (MSE's) of guess 1, guess 2 and the average of both guesses in the immediate condition and the delayed condition. Error bars represent standard errors.*

estimate is not significantly smaller than $d_{33\%}$. Using G*Power 3.1, we calculated that $d_{33\%} = .10$ in the immediate condition and $d_{33\%} = .12$ in the delayed condition, based on the sample sizes from Vul and Pashler (2008). Clearly, our effect size estimates of both guesses in both conditions are larger than $d_{33\%}$, so against this criterion also, the current study is a successful replication of the crowd within effect in both the immediate and the delayed condition.

Following Vul and Pashler (2008), we performed two additional analyses. First, both in the immediate and in the delayed condition, second guesses were less accurate than first guesses. In the immediate condition, the mean MSE of guess 1 ($M = 589, SD = 336$) was smaller than the mean MSE of guess 2 ($M = 615, SD = 351$), $t(470) = -2.25, p = .025$, $\hat{d_z} = -.10$, 95% CI $= [-.19, -.01]$. Likewise, in the delayed condition, the mean MSE of guess 1 ($M = 517, SD = 288$) was smaller than the mean MSE of guess 2 ($M = 589, SD = 327$), $t(139) = -2.91, p = .004$, $\hat{d_z} = -.25$, 95% CI $= [-.41, -.08]$. These results reassure that the accuracy gain of averaging could not be attributed to participants looking up the answers between guesses.

This is also confirmed by the scatter plots with the marginal histograms of the MSE's of guess 1 and guess 2 in both conditions (see Figure 1.2). As noted by Vul (n.d.), if participants were looking up the answers, there should be a peak in the error histograms at the value that can be expected when people know the right answer, i.e. error = 0. Clearly, this is not the case in Figure 1.2.

Second, unlike in Vul and Pashler (2008), the accuracy gain of averaging both guesses compared to guess 1 was not significantly larger in the delayed condition than in the immediate condition.[9] The mean difference between the MSE of the average and the MSE of guess 1 was not significantly larger in the delayed condition ($M = 50, SD = 147$) than in the immediate condition ($M = 48, SD = 119$), $t(609) = 0.18, p = .858$, $\hat{d} = .02$, 95% CI = $[-.17, .21]$.

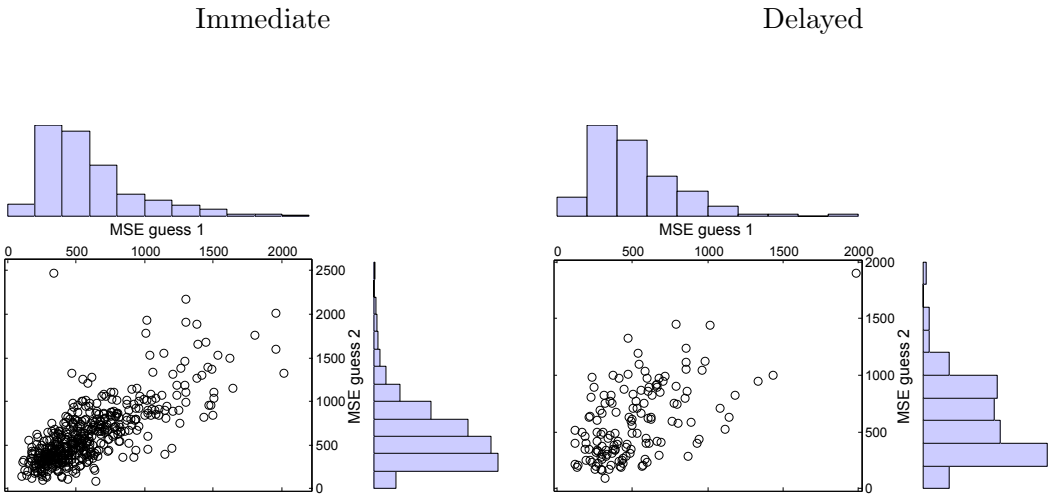


Figure 1.2: *Mean squared errors (MSE's) of guess 1 and guess 2 in the immediate (Panel A) and in the delayed (Panel B) condition.*

[9] Contacting the authors concerning the experiment led them to uncover that their data show a slightly stronger evidence for this accuracy gain effect between both conditions than what had been reported in the article originally. The reported t-statistic, $t(426) = 2.12, p < .05$, was smaller than the correct t-value, $t(426) = 2.68, p = .008$.

Post-hoc Analyses

Since we were surprised by the non-significant difference between the immediate and the delayed condition in the accuracy gain of averaging guesses compared to guess 1, we also tested the difference between both conditions by comparing the average guess to guess 2. Unlike our comparison with guess 1, the mean difference between the MSE of the average and the MSE of guess 2 was significantly larger in the delayed condition ($M = 121, SD = 169$) than in the immediate condition ($M = 73, SD = 155$), $t(609) = 3.14, p = .002$, $\hat{d} = .30$, 95% CI = $[.11, .49]$. As this comparison was not reported in the original paper, we performed the same analysis on the raw data of the original study, which were provided to us by Edward Vul. Unlike in the current study, the mean difference between the MSE of the average and the MSE of guess 2 was not significantly larger in the delayed condition ($M = 164, SD = 218$) than in the immediate condition ($M = 131, SD = 211$), $t(426) = 1.56, p = .12$, $\hat{d} = .15$.

In sum, the evidence for the difference in magnitude of the crowd within effect between the immediate condition and the delayed condition is mixed. Whereas the original study by Vul and Pashler (2008) yields a significant difference between both conditions when the average guess is compared to guess 1, but not when it is compared to guess 2, the present study yields the opposite pattern: When the average guess is compared to guess 2, there is a significant difference between both conditions, but when it is compared to guess 1, the difference is not significant. Thus, in both studies, introducing a three-week delay increased the benefit of averaging compared to one of both guesses only. However, it is important to keep in mind the exploratory nature of these analyses, as we did not use a power analysis to determine the sample size for testing these effects and we did not a priori specify the comparison of the average with guess 2.

Bayes Factors for Confirmatory and Post-hoc Analyses

The Bayesian alternative to null hypothesis significance testing is calculating the Bayes factor (BF), which quantifies the evidence of the null hypothesis relative to the alternative hypothesis. In contrast to a p-value, the BF can provide evidence both in favor or against the null hypothesis. Therefore, in addition to p-values, we calculated BF's for all tests, using a web-based Bayes factor calculator.[10] Table 1.5 shows the BF's for all tests of the confirmatory and post-hoc analyses, together with the t-statistics, sample sizes and p-values. All BF's show qualitatively identical results as the null hypothesis significance tests (i.e., tests with $p < .05$ have BF < 1, indicating evidence for the alternative hypothesis, whereas tests with $p > .05$ have $BF > 1$, indicating evidence for the null hypothesis), except for the comparison between the error of guess 1 and guess 2 in the immediate condition. The p-value for this latter test is .025, suggesting the difference is significant, whereas the BF is 2.208, indicating anecdotal evidence for the null hypothesis of no difference. However, both results are consistent with participants not having looked up the answers between guesses.

Discussion

Our replication attempt of the crowd within effect supports the original finding by Vul and Pashler (2008) that averaging two guesses within one person provides a more accurate answer than either guess alone. This effect was found when the second guess was made immediately after the first guess (immediate condition), as well as when the second guess was made three weeks later (delayed condition). These results were evaluated as success-

[10] The Bayes factor calculator, available at `http://pcl.missouri.edu/bayesfactor`, requires input of the t-statistic and the sample size only. Following Rouder et al. (2009)'s recommendations, we report the JZS Bayes factor.

Table 1.5: *JZS Bayes factors (BF, with scale $r = 1$) in favor of the null hypothesis of no difference for all tests.*

t	n	p	BF
8.69	471	$< .001$	$1.8\ 10^{-14}$
10.26	471	$< .001$	$7.1\ 10^{-20}$
4.02	140	$< .001$	.007
8.48	140	$< .001$	$4.7\ 10^{-12}$
-2.25	471	.025	2.208
-2.91	140	.004	0.252
0.18	471 (n_1) & 140 (n_2)	.858	12.931
3.14	471 (n_1) & 140 (n_2)	.002	0.107

Note: The first seven tests are from the confirmatory analyses, the last test is from the post-hoc analyses.

ful replications against two different replication evaluation standards: The traditional p-value approach on the one hand, and the recently proposed detectability approach on the other hand.

The three-week delay between the two guesses improved the accuracy gain of averaging compared to guess 2, but not compared to guess 1. These results are comparable to those in Vul and Pashler (2008), where an increase in accuracy gain was also observed with the comparison of one of both guesses only. Thus, it seems that more research is needed to investigate whether a temporal separation between guesses can boost the crowd within effect.

Appendices

Appendix A: Pre-registered Matlab Script for Confirmatory Analyses

```
% This code is based on code written by Edward Vul,
% except for the computation of the confidence intervals ... for the
% effect sizes, which is based on code provided by ... Simonsohn  (2013).
data;
% The data are 2 nx16 matrices (the first matrix with data from
% immediate condition; the second matrix with data from the ... the delayed
% condition), where n is the sample size of the ... corresponding condition.
% Each row corresponds to the answers from one participant. ... The first 8
% columns correspond to the first guesses; the final 8 ... colums correspond to
% the second guesses.

answers = [6.3 43.3 32.3 13.4 53.6 54.8 26.4 22.4];
% These answers are derived from The World Factbook ... (Central Intelligence
% Agency, 2013)

%% sets are 1: immediate, 2: delayed

for set = [1:2]
    [si sj] = size(data{set});
    grp(set).n=si;
    % guess 1
    grp(set).g{1} = data{set}(:,[1:8]);
```

```
23      % guess 2
24      grp(set).g{2} = data{set}(:,[9:16]);
25      % average of guesess
26      grp(set).g{3} = (grp(set).g{1}+grp(set).g{2})./2;
27
28      for g = [1:3]
29          %[n k] = size(grp(set).g{1});
30          bigans = repmat(answers, grp(set).n, 1);
31          % mean squared error
32          grp(set).d{g} = mean((grp(set).g{g}-bigans).^2,2);
33          grp(set).mumse(g) = mean(grp(set).d{g});
34          grp(set).semse(g) = ...
                std(grp(set).d{g})./sqrt(grp(set).n);
35      end
36  end
37
38  %% graph of mean MSE guess 1, guess2, average guess
39
40  figure();
41  barweb([grp(1).mumse; grp(2).mumse], [grp(1).semse; ...
        grp(2).semse]);
42  ylim([400 700]);
43
44  %% comparisons between guess and average
45
46  for set = [1:2]
47      for g = [1:2]
48      % guess 1 or guess 2 compared to average.
49      % ttest
50      [h p ci stats] = ttest(grp(set).d{g} - grp(set).d{3});
51      grp(set).t{g} = stats.tstat;
52      grp(set).p{g} = p;
53      % effect size
54      grp(set).dz{g} = grp(set).t{g}/sqrt(grp(set).n);
55      % confidence interval (see Cumming and Finch (2001, pp. ...
            544-545))
```

```
56      alpha=.05;
57      df = grp(set).n-1;
58      tnonct = inline('nctcdf(x,df,delta) - pr');
59      ncp_low = fzero(@(delta) tnonct(delta, df, 1-alpha/2, ...
            grp(set).t{g}), [-20,20]);
60      ncp_high = fzero(@(delta) tnonct(delta, df, alpha/2, ...
            grp(set).t{g}), [-20,20]);
61      grp(set).dzlow{g} = ncp_low/sqrt(grp(set).n);
62      grp(set).dzhigh{g} = ncp_high/sqrt(grp(set).n);
63      end
64
65      % guess 1 compared to guess 2.
66      [h3 p3 ci3 stats3] = ttest(grp(set).d{1} - grp(set).d{2});
67  end
68
69
70  %% comparison between magnitude of averaging benefit over ...
        guess 1.
71  [h4 p4 ci4 stats4] = ttest2(grp(2).d{1} - grp(2).d{3}, ...
        grp(1).d{1} - grp(1).d{3});
```

Appendix B: Used Matlab Script for Confirmatory Analyses

```
% This is the code used for the confirmatory analyses.
% Apart from minor syntactic clean-up, this code differs ... from the
% pre-registered code in Appendix A in that we added code for
% 1) descriptive statistics
% 2) the scatter histogram plots in Figure 2
% 3) some effect sizes
% 4) some confidence intervals for effect sizes.
% Further, the appearance of the bar chart (Figure 1) is ... changed.

% The data file (data_steegenetal2014.mat) is made from ... several .txt files,
% using the code from steegenetal2014_postprocessing.m

% This code is based on code written by Ed Vul,
% except for the computation of the confidence intervals ... for the
% effect sizes, which is based on code provided by ... Simonsohn  (2013).
clear all
close all
load data_steegenetal2014;
% The data are 2 nx16 matrices (the first matrix with data from
% immediate condition; the second matrix with data from the ... the delayed
% condition), where n is the sample size of the ... corresponding condition.
% Each row corresponds to the answers from one participant. ... The first 8
% columns correspond to the first guesses; the final 8 ... colums correspond to
```

```
% the second guesses.

% The code relies on barweb.m, which is available on ...
    matlabcentral

% Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. ...
    (2014). Measuring the crowd within again: A ...
    pre-registered replication study. Frontiers in Psychology

answers = [6.3 43.3 32.3 13.4 53.6 54.8 26.4 22.4];
% These answers are derived from The World Factbook ...
    (Central Intelligence
% Agency, 2013)

% answers = [6.3 44.4 30.3 10.5 58 72.4 18.9 20.3]; these ...
    are the answers
% used in Vul & Pashler (2008) (see first colomn Table 3)

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% compute MSE, descriptive statistics and plot data
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

%% sets are 1: immediate, 2: delayed

for set = [1:2]
    [si sj] = size(data{set});
    grp(set).n=si;
    % guess 1
    grp(set).g{1} = data{set}(:,[1:8]);
    % guess 2
    grp(set).g{2} = data{set}(:,[9:16]);
    % average of guesses
    grp(set).g{3} = (grp(set).g{1}+grp(set).g{2})./2;

    for g = [1:3]
        bigans = repmat(answers, grp(set).n, 1);
```

```
55          % mean squared error
56          grp(set).d{g} = nanmean((grp(set).g{g}-bigans).^2,2);
57          grp(set).mumse(g) = nanmean(grp(set).d{g});
58          grp(set).semse(g) = ...
                nanstd(grp(set).d{g})./sqrt(grp(set).n);
59      end
60  end
61
62  % descriptive statistics
63  for set = [1:2]
64  grp(set).sdmse = grp(set).semse.*sqrt(grp(set).n);
65      for g=[1:2]
66          grp(set).corr{g}=corr(grp(set).d{g}, grp(set).d{3});
67          grp(set).mudiff{g} = mean(grp(set).d{g} - ...
                grp(set).d{3});
68          grp(set).sddiff{g} = std(grp(set).d{g} - ...
                grp(set).d{3});
69      end
70  end
71
72  %% graph of mean MSE guess 1, guess2, average guess
73  figure();
74  barweb([grp(1).mumse; grp(2).mumse], [grp(1).semse; ...
        grp(2).semse], [],{'Immediate'; 'Delayed '},[],[],'Mean ...
        Squared Error',[0 0 0; 1 1 1; .5 .5 .5],[],{'Guess 1'; ...
        'Guess 2'; 'Average'},'WestEast');
75  ylim([0 850]);
76  legend({'Guess 1'; 'Guess 2'; ...
        'Average'},'Location','Northeast')
77
78  % scatter histogram plots
79  figure();
80  scatterhist(grp(1).d{1},grp(1).d{2}, 'Location', ...
        'NorthEast', 'Direction', 'Out', 'Color', 'k')
81  xlabel('MSE guess 1', 'FontSize', 12)
82  ylabel('MSE guess 2', 'FontSize', 12)
```

```
figure
scatterhist(grp(2).d{1}, grp(2).d{2}, 'Location', ...
    'NorthEast', 'Direction', 'Out', 'Color', 'k')
xlabel('MSE guess 1', 'FontSize', 12)
ylabel('MSE guess 2', 'FontSize', 12)


%%%%%%%%%%%%%%%%%%%
% inferential tests
%%%%%%%%%%%%%%%%%%%

alpha=.05;

%% comparisons between guess and average and between guess ...
    1 and guess 2

for set = [1:2]
    % guess 1 or guess 2 compared to average.
    for g = [1:2]
    % ttest
    [h p ci stats] = ttest(grp(set).d{g} - grp(set).d{3});
    grp(set).t{g} = stats.tstat;
    grp(set).df{g} = stats.df;
    grp(set).p{g} = p;
    % effect size
    grp(set).dz{g} = grp(set).t{g}/sqrt(grp(set).n);
    % confidence interval (see Cumming and Finch (2001, pp. ...
        549-550))
    df = grp(set).df{g};
    tnonct = inline('nctcdf(x,df,delta) - pr');
    ncp_low = fzero(@(delta) tnonct(delta, df, 1-alpha/2, ...
        grp(set).t{g}), [-20,20]);
    ncp_high = fzero(@(delta) tnonct(delta, df, alpha/2, ...
        grp(set).t{g}), [-20,20]);
    grp(set).dzlow{g} = ncp_low/sqrt(grp(set).n);
    grp(set).dzhigh{g} = ncp_high/sqrt(grp(set).n);
```

```
114     end
115
116     % guess 1 compared to guess 2.
117     % ttest
118     [h p ci stats] = ttest(grp(set).d{1} - grp(set).d{2});
119     grp(set).t{3} = stats.tstat;
120     grp(set).df{3} = stats.df;
121     grp(set).p{3} = p;
122     % effect size
123     grp(set).dz{3} = grp(set).t{3}/sqrt(grp(set).n);
124     % confidence interval
125     df = grp(set).df{3};
126     tnonct = inline('nctcdf(x,df,delta) - pr');
127     ncp_low = fzero(@(delta) tnonct(delta, df, 1-alpha/2, ...
            grp(set).t{3}), [-20,20]);
128     ncp_high = fzero(@(delta) tnonct(delta, df, alpha/2, ...
            grp(set).t{3}), [-20,20]);
129     grp(set).dzlow{3} = ncp_low/sqrt(grp(set).n);
130     grp(set).dzhigh{3} = ncp_high/sqrt(grp(set).n);
131 end
132
133
134 %% comparison of magnitude of averaging benefit over guess ...
        1 between immediate and delayed condtion.
135 % ttest
136 [h p ci stats] = ttest2(grp(2).d{1} - grp(2).d{3}, ...
        grp(1).d{1} - grp(1).d{3});
137 grpcmp.t{1} = stats.tstat;
138 grpcmp.df{1} = stats.df;
139 grpcmp.p{1} = p;
140 % effect size (cohen's standardized mean difference d for ...
        independent groups; see Cumming and Finch (2001, pp. 567)
141 grpcmp.d{1} = grpcmp.t{1}.*sqrt((1/grp(1).n)+(1/grp(2).n));
142 % confidence interval (see Cumming and Finch (2001, pp. 567))
143 df = grpcmp.df{1};
144 tnonct = inline('nctcdf(x,df,delta) - pr');
```

```
145 ncp_low = fzero(@(delta) tnonct(delta, df, 1-alpha/2, ...
        grpcmp.t{1}), [-20,20]);
146 ncp_high = fzero(@(delta) tnonct(delta, df, alpha/2, ...
        grpcmp.t{1}), [-20,20]);
147 grpcmp.dlow{1} = ncp_low.*sqrt((1/grp(1).n)+(1/grp(2).n));
148 grpcmp.dhigh{1} = ncp_high.*sqrt((1/grp(1).n)+(1/grp(2).n));
```

Chapter 2

Increasing transparency through a multiverse analysis

Introduction

Psychology has been stirred by dramatic revelations of questionable research practices (John et al., 2012), implausible findings (Wagenmakers et al., 2011), and low reproducibility (Open Science Collaboration, 2015; Yong, 2012). The resulting crisis of confidence has led to a wide array of recommendations for improving research practices. Commonly cited advice includes replication, high power, co-piloting, adjusting the alpha level, focusing on estimation rather than on testing, and adopting Bayesian statistics (e.g., Asendorpf et al., 2013; Bakker, van Dijk, & Wicherts, 2012; Johnson, 2013; Wagenmakers et al., 2011). A major class of recommendations involves a call for increased transparency in reporting, including preregistration of hypotheses and analyses, clearly distinguishing between confirmatory and exploratory findings, disclosing all conditions and measures, sharing data, and sharing research materials (e.g., Chambers, 2013; LeBel, Campbell, & Loving, in press; Morey et al., 2016; Nosek & Bar-Anan, 2012; Nosek et al., 2015; Simmons et al., 2012; Wagenmakers et al., 2012). In this chapter, we use a worked example to suggest that research transparency can further be

increased by performing what we term a *multiverse analysis*.

A multiverse analysis starts from the observation that data used in an analysis are usually not just passively recorded in an experiment or an observational study. Rather, data are to a certain extent actively *constructed*. Data construction occurs when the raw data are converted into a form ready for analysis. When preparing their data for analysis, researchers often take several processing steps, such as discretization of variables into categories, combination of variables, transformation of variables, data exclusion, and so on. These processing steps typically come with many researcher degrees of freedom (Simmons et al., 2011), as there are often several options in each step. As a result, raw data do not uniquely give rise to a single data set for analysis, but rather to multiple alternatively processed data sets, depending on the specific combination of choices—a *many worlds* or *multiverse* of data sets. As each data set in this data multiverse can lead to a different statistical result, the data multiverse directly implies a multiverse of statistical results.

Researchers often select a single (or a few) data-processing choices and then present this as the only analysis that ever would have been done. This practice of selective reporting would not be problematic if the single data set under consideration is processed based on sound and justifiable choices. However, choosing among the possibilities during data processing is often arbitrary, and justifications for the choices are typically lacking. For example, partitioning a variable into two or more discrete categories often involves an arbitrary split point, there can be various reasonable combinations or transformations of variables, and there are different sensible guidelines to determine which data points to exclude. This multiplicity of reasonable processing steps gives rise to a multiverse of reasonable data sets, which directly implies that there are several reasonable statistical results. Any arbitrariness that is present in the data construction is inherited by the

statistical result.

When privileging a single arbitrary data set from the multiverse of possible data sets, the multiverse of statistical results is ignored. The inevitable arbitrariness in the data and the sensitivity of the result is hidden to the readers, which makes the interpretation of the single result hard at best and impossible at worst. In the light of this problem of selective reporting, we propose to use a *multiverse analysis* as an alternative to a single data set analysis. Such a multiverse analysis has two goals: It enhances transparency by providing a detailed picture of the robustness or fragility of statistical results, and it helps identifying the key choices that conclusions hinge on.

A multiverse analysis involves performing the analysis of interest across the whole set of data sets that arise from different reasonable choices for data processing. It can be seen as a systematic and organized extension of *outlier analysis* (see, e.g., Ramsey & Schafer, 2012; Simmons et al., 2011), which involves examining the robustness of one's conclusions with and without the elimination of outlying observations. A multiverse analysis displays the stability or robustness of an effect, not only across different options for exclusion criteria, but across different options for *all* steps in data processing. It is closely related to the idea of a *garden of forking paths* in data analysis (Gelman & Loken, 2014), which highlights that the one-to-many mapping from scientific theories to statistical hypotheses typically leads to an implicit, potential multiple comparison problem. The multiverse analysis focuses on one particular aspect of this multiple comparison issue, related to data processing.

In the remainder of this chapter, we demonstrate a multiverse analysis using data from recently published research. We first describe the results of an analysis focusing on a single constructed data set only. Next, we describe a multiverse analysis based on the same raw data and highlight how the multiverse analysis reveals the impact of arbitrary processing choices on the

statistical results.

Demonstration

Our demonstration of a multiverse analysis focuses on data collected by Durante, Rae, and Griskevicius (2013). These authors conducted two studies investigating the effect of fertility on religiosity and political attitudes. We selected this paper simply to illustrate how a multiverse analysis can help researchers better understand the extent to which their results depend on various data processing choices. First, we describe the raw data that were collected in both studies. Next, we describe the single data set analysis reported by Durante et al. (2013). Finally, we show what these authors could have found, had they performed a multiverse analysis of their data rather than the single data set analysis. A more detailed description of the raw and processed data is provided in the Appendix.

Data collection

A total of 275 women participated in Study 1. Each participant was asked to answer three religiosity items using a 9-point scale. Further, each participant was asked to indicate the typical length of her menstrual cycle, the start date of her last menstrual period, and the start date of her previous menstrual period. In addition, each woman indicated how sure she was about these two start dates, using a 9-point scale. Finally, each woman was asked to indicate her current romantic relationship status, with the following four response options: (1) *not dating/romantically involved with anyone*, (2) *dating or involved with only one partner*, (3) *engaged or living with my partner*, and (4) *married*.

Quite laudably, Durante et al. (2013) performed a second study to replicate the findings in Study 1, and to extend them to political attitudes. In

Study 2, 502 women participated. The main difference with Study 1 was that participants were also asked to answer five items assessing fiscal political attitudes, five items assessing social political attitudes (using a 7-point scale for these ten items), one item assessing their voting preference (Mitt Romney or Barack Obama), and one item assessing their campaign donation preference (Mitt Romney or Barack Obama). Another difference with Study 1 was that participants also indicated the expected start date of their next menstrual period.

Single data set analysis

The data collected in the procedure described above are not ready for analysis yet. Preparing the data set for analysis requires several processing steps and decisions. We describe the different data processing steps taken by Durante et al. (2013) to construct a single data set for each study, and the main results and conclusions that follow from this data set. The results of these single data set analyses are identical to the ones reported by Durante et al. (2013).

Constructing the single data set

In order to construct a single data set ready for analysis, the following data processing steps are taken.

Religiosity The three religiosity items are averaged to create a religiosity score.

Fiscal and social political attitudes The five fiscal political attitudes items are averaged to create a fiscal political attitudes score, and the five social political attitudes items are averaged to create a social political attitudes score.

Fertility Participants are classified in a *high* or *low* fertility group based on their cycle day. Participants with cycle days ranging from 7 to 14 are assigned to the high fertility group, whereas participants with cycle days ranging from 17 to 25 are assigned to the low fertility group. A woman's cycle day is based on the number of days before next menstrual onset, which in turn is based on cycle length, which is computed as the difference between the start date of the woman's last menstrual period and the start date of the woman's previous menstrual period.

Relationship status Participants are assigned to a *single* versus *committed relationship* group. Women who selected response Option (1) or (2) on the relationship status item are assigned to the group of single women, whereas women who selected response Option (3) or (4) are assigned to the group of women in committed relationships.

Exclusion criteria The assignment of the participants to a high or low fertility group automatically excludes women whose cycle days are not in the high or low fertility range. Beyond this exclusion, no other participants are excluded.

Deriving the single statistical result

Based on this single data set, the effect of fertility on religiosity and political attitudes is examined, with relationship status as an interacting variable. For religiosity, an ANOVA reveals a Fertility × Relationship status interaction, in both studies — $F(1, 159) = 6.46$, $p = 0.012$, in Study 1; $F(1, 299) = 8.21$, $p = 0.004$ Study 2 —, indicating that single women reported less religiosity if they were in the high-fertility group than if they were in the low-fertility group, whereas women in relationships reported more religiosity if they were in the high-fertility group than in the low-fertility group. Regarding fiscal political attitudes, an ANOVA reveals no significant effects

of fertility status. Regarding social political attitudes, a Fertility × Relationship status interaction is found, $F(1, 299) = 12.26, p = .001$, indicating that single women reported less socially conservative attitudes if they were in the high-fertility group than if they were in the low-fertility group, whereas women in relationships showed the opposite pattern. Finally, logistic regression reveals a significant Fertility × Relationship status interaction both for voting preferences, $b = -1.62$, $Wald(1) = 8.35$, $p = .004$, and donation preferences, $b = -1.71$, $Wald(1) = 9.30$, $p = .002$, indicating that single women were more likely to vote and donate for Obama if they were in the high-fertility group than if they were in the low-fertility group, whereas women in relationships were more likely to vote and donate for Romney if they were in the high-fertility group than if they were in the low-fertility group.

Multiverse analysis

The different data processing steps in the single data set analysis are far from the only reasonable ones (see also Harris, Pashler, & Mickes, 2014). This means that the data set used in the single data set analysis corresponds to just a single data set in a much larger multiverse of data sets. More importantly, this also means that the statistical result based on the single data set reflects only one possible outcome in a multiverse of possible outcomes. Without knowing which other statistical results could have reasonably be observed, it is impossible to evaluate the robustness of the finding. Transparency could be increased by performing, for each research question, the same analysis for *all* possible data sets, defined by the reasonable choices for data processing. This is the multiverse analysis.

We will first construct the multiverse of data sets, which consists of all data sets that could be obtained by combining different reasonable data processing choices. Then, we analyze each data set in this data multiverse separately, leading to the multiverse of statistical results. In this multiverse

analysis, we consider choices in data processing that Durante et al. (2013) might themselves have considered had they performed a multiverse analysis rather than a single data set analysis. To increase the likelihood that these authors would have considered these choices reasonable, the different processing choices we use are based on previously published studies by Durante and her collaborators, where possible. In the same spirit, we followed Durante et al. (2013) in dichotomizing the relationship status and fertility variables, although the practice dichotomization is not without criticism (e.g., MacCallum, Zhang, Preacher, & Rucker, 2002).[1] Further, the vicarious character of our multiverse analysis implies that for the construction of the multiverse of results, we will adopt the statistical analyses that were used by Durante et al. (2013), including the focus on p-values and the adoption of .05 as the significance level. We stress that this is only a hypothetical illustration of a multiverse analysis. Our multiverse is only a subset of a larger multiverse of possible data processing choices, and we can not rule out that Durante et al.'s (2013) actual multiverse might have been different.

Constructing the data multiverse

The first step involves listing the different reasonable choices during each step of data processing. Box 1 summarizes five arbitrary choices in data processing, both in Study 1 and 2, and the different reasonable options we will consider for each arbitrary choice. Option (a) always corresponds to the processing choice made by Durante et al. (2013), while the remaining options correspond to alternative choices they could have reasonably made. In the following sections, we describe the alternative options in detail.

[1] For one of the six analyses of interest, Durante et al. (2013) report an additional analysis that uses a continuous measure of fertility, conception probability, rather than the dichotomized one, maybe inspired by these criticisms (see also Gangestad et al., 2016). However, since the majority of their analyses uses a dichotomized assessment of fertility, we will do so here as well.

Box 1. Processing choices

1. Assessment of fertility (F) (high vs low)
   (a) F1: high = cycle days 7–14; low = cycle days 17–25
   (b) F2: high = cycle days 6–14; low = cycle days 17–27
   (c) F3: high = cycle days 9–17; low = cycle days 18–25
   (d) F4: high = cycle days 8–14; low = cycle days 1–7 and 15–28
   (e) F5: high = cycle days 9–17; low = cycle days 1–8 and 18–28
2. Next menstrual onset (NMO)
   (a) NMO1: reported start date previous menstrual onset + computed cycle length
   (b) NMO2: reported start date previous menstrual onset + reported cycle length
   (c) NMO3: reported estimate of next menstrual onset
3. Assessment of relationship status (R) (single vs relationship)
   (a) R1: single = response options 1 and 2; relationship = response options 3 and 4
   (b) R2: single = response option 1; relationship = response options 2, 3, and 4
   (c) R3: single = response option 1; relationship = response options 3 and 4
4. Exclusion of women based on cycle length (ECL)
   (a) ECL1: no exclusion based on cycle length
   (b) ECL2: exclusion of participants with computed cycle length < 25 or > 35 days
   (c) ECL3: exclusion of participants with reported cycle length < 25 or > 35 days
5. Exclusion of women based on certainty ratings of start dates of two previous menstrual periods (EC)
   (a) EC1: no exclusion based on certainty ratings
   (b) EC2: exclusion of participants who are not certain about at least one start date (i.e., sure < 6)

Fertility First, the classification of women into a high or low fertility group based on cycle day can be done using several reasonable alternatives: assigning women with cycle days 6–14 to the high fertility group and women with cycle days 17–27 to the low fertility group (e.g., Durante, Griskevicius, Hill, Perilloux, & Li, 2011), days 9–17 for high fertility and 18–25 for low fertility (Durante, Griskevicius, Simpson, Cantú, & Li, 2012), days 8–14 for high fertility and 1–7 and 15–28 for low fertility (Durante, Griskevicius, Cantú, & Simpson, 2014), and days 9–17 for high fertility and 1–8 and 18–28 for low fertility (Durante & Arsena, 2015).

Second, there are different reasonable alternatives for estimating a woman's next menstrual onset, which is an intermediate step in determining cycle day. A reasonable way to estimate next menstrual onset involves using the women's reported estimate of their typical cycle length (e.g., Thornhill & Gangestad, 1999). Another reasonable strategy for determining the onset of the next period involves using the self-reported expected start date of the next menstrual period (e.g., Haselton & Miller, 2006).[2]

Relationship status There are at least two reasonable alternative options to the dichotomization of women's relationship status, stemming from the ambiguous nature of response Option 2 (*dating or involved with only one partner*). This option can cover both single women (*dating*) or women in relationships (*involved with only one partner*). Thus, women who select this response could reasonably be classified as being either in committed relationships or as being single. A third option involves discarding participants who select this ambiguous response option, and only classifying participants selecting Option 1 as single women and participants selecting Option 3 or 4 as women in relationships.

[2]The fact that typical cycle length and the expected start date of the next period were collected by Durante et al. (2013) suggests that they considered this option at least somewhat reasonable.

Exclusion criteria First, it is not unreasonable to exclude participants with irregular cycle lengths. This could amount to only including women with cycle lengths 25 to 35 (e.g., Durante et al., 2012). This exclusion criterion can be instantiated in two reasonable ways, using either a woman's computed cycle length or a woman's self-reported typical cycle length.

Second, another justifiable exclusion criterion concerns women's reported certainty ratings of the start dates of their last two menstrual periods. It is reasonable to exclude participants who were not sufficiently confident about their report, and to consider only data from participants with a certainty rating above the midpoint for both dates (e.g., Durante, Arsena, & Griskevicius, 2014).

Based on this tabulation of choices, the multiverse of data sets is constructed by considering all combinations of reasonable choices in data processing, and deriving a data set for each of the different choice combinations. In Study 1, there are $5 \times 2 \times 3 \times 3 \times 2 = 180$ choice combinations (see Box 1; NMO3, the estimation of next menstrual onset based on the reported estimate, could not be applied to Study 1, as the expected start date of the next menstrual period was not collected in this study). Some of the choice combinations are inconsistent: When participants are excluded based on reported or computed cycle length, we do not consider next menstrual onset based on computed or reported cycle length, respectively. After excluding these inconsistent combinations, we are left with $180 - 2 \times (5 \times 1 \times 3 \times 1 \times 2) = 120$ choice combinations. Similarly, in Study 2, there are $5 \times 3 \times 3 \times 3 \times 2 = 270$ choice combinations, but after excluding inconsistent combinations, $270 - 2 \times (5 \times 1 \times 3 \times 1 \times 2) = 210$ choice combinations remain.

Deriving the multiverse of statistical results

After constructing the data multiverse, the analysis of interest (in this case, an ANOVA or a logistic regression) is performed across all the alternatively constructed data sets.[3] The results are shown in Panels A-F of Figure 2.1, each showing a histogram of the p-values of the Fertility $\times$ Relationship interaction effect.

For two variables —religiosity in Study 1 (Panel A) and fiscal political attitudes (Panel C) — the multiverse analysis reveals a near-uniform distribution, indicating that the p-value for the interaction effect between fertility and relationship varies widely across the multiverse. For religiosity, seven out of the 120 choice combinations lead to a significant interaction effect, whereas the remaining 94% lead to p-values ranging from .05 to 1.0. For fiscal political attitudes, 8% of the 210 choice combinations lead to a significant interaction ($p < .05$), whereas the remaining choice combinations lead to p-values across the entire range from .05 to 1.0.

For the remaining four variables, roughly half of the choice combinations lead to a significant interaction effect. In particular, for religiosity in Study 2 (Panel B), 88 out of the 210 choice combinations (42%) lead to a p-value smaller than .05. Regarding social political attitudes (Panel D), 49% of the p-values is smaller than .05. Finally, 46% and 57% of the p-values are smaller than .05 for voting (Panel E) and donation (Panel F) preferences, respectively. In these cases, it is informative to display the multiverse in greater detail by showing which constellation of choices corresponds to which statistical result. This allows to identify the key choices in data processing that are most consequential in the fluctuation of the statistical results.

Such a closer inspection is provided in Figure 2.2, showing a grid of p-values for each of these four variables. In each panel, the cells show the

[3]Due to coding errors in the data, there were some missing data (see Appendix). In our analyses, incomplete cases are discarded.

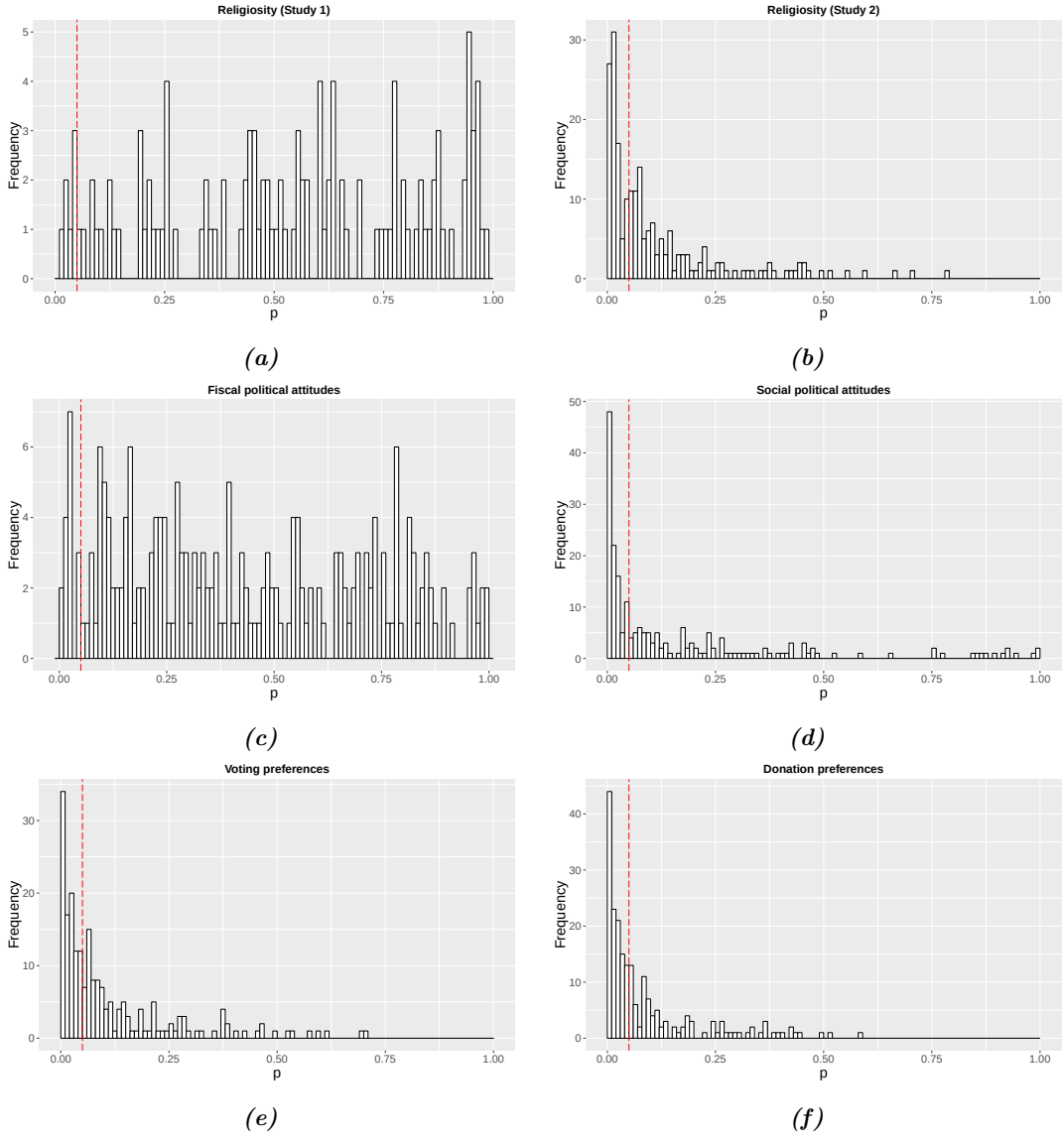


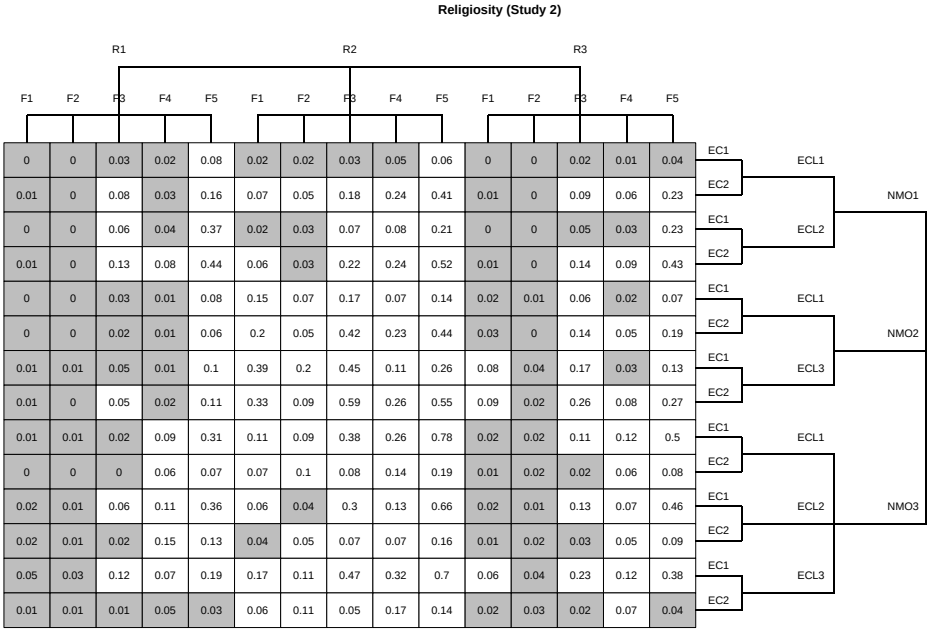
Figure 2.1: *Histogram of p-values for the Fertility x Relationship status interaction effect on religiosity for the multiverse of 120 data sets in Study 1 and 210 data sets in Study 2 (Panels A and B), on fiscal and social political attitudes for the multiverse of 210 data sets in Study 2 (Panels C and D), and on voting and donation preferences for the multiverse of 210 data sets in Study 2 (Panels E and F). The dashed line indicates* $p = .05$.

different p-values that can be obtained across all choice combinations for data processing. Depending on whether the p-value is smaller or larger than the α-level, the cells are colored gray or white, respectively.

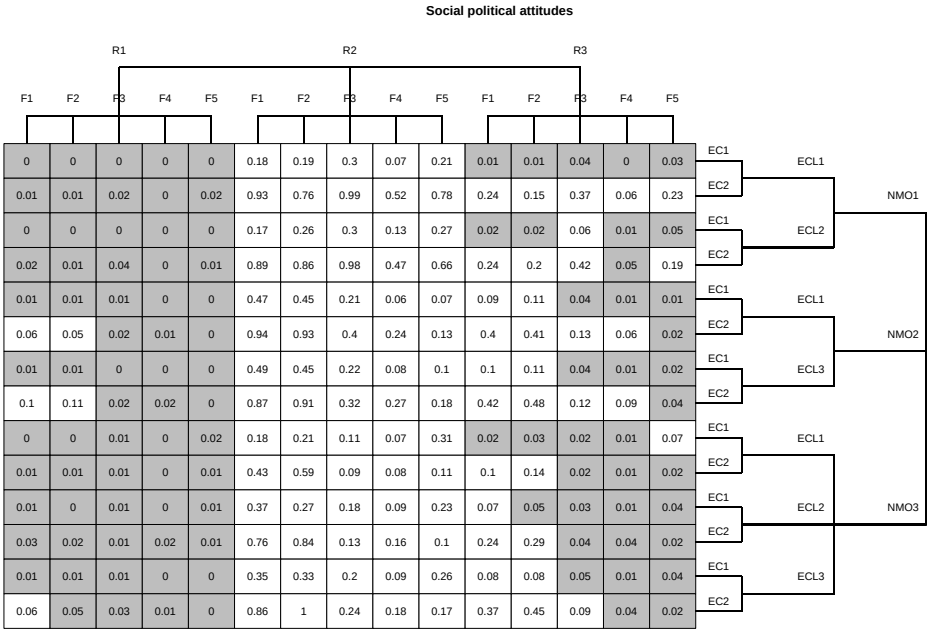
For religiosity in Study 2 (Panel A), most data sets constructed under the second option for relationship assessment (R2) yield a non-significant interaction effect. The first and third options (R1 and R3) consistently lead to a significant interaction effect in combination with the first and second option for fertility assessment (F1 and F2) and to a non-significant interaction effect in combination with F5, whereas data sets constructed under R1 or R3 in combination with F3 or F4 lead to more fluctuating conclusions, depending on the other choices for data processing. The different exclusion criteria and cycle day estimation options do not seem to have a large impact on fluctuation in the statistical conclusion. For social political attitudes (Panel B), the statistical conclusion is highly robust for the first and second option for relationship status assessment (significant for R1 and non-significant for R2). Using the third option for relationship status assessment (R3) leads to more fluctuation, depending on the choices for the other processing steps. Finally, for voting and donation preferences (Panels C and D, respectively), it is hard to extract a consistent pattern of fluctuation across the different choice combinations. It seems that all arbitrary choices for data processing can have an impact on whether the obtained data set will lead to a significant or a non-significant outcome.

Discussion

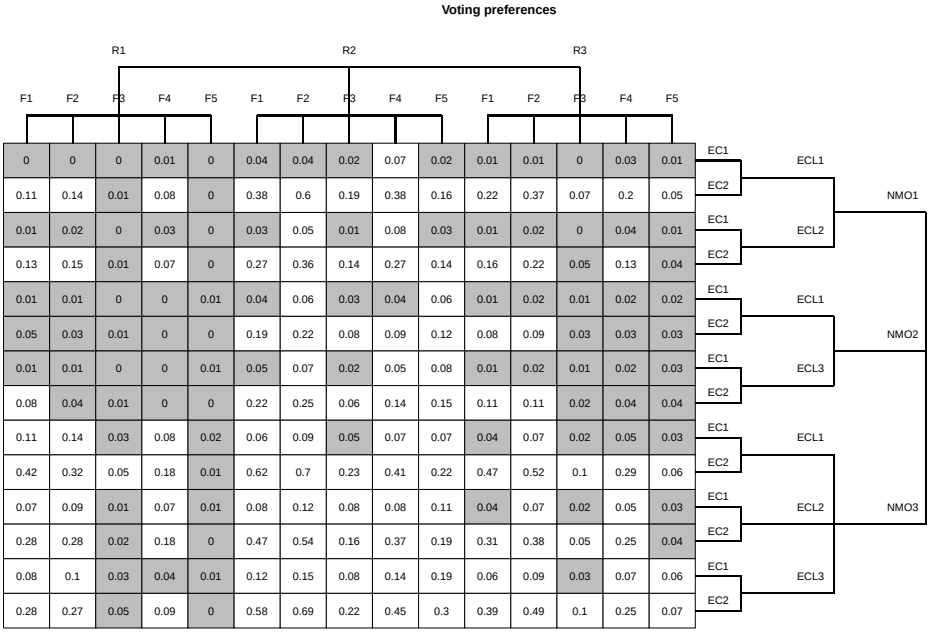
Converting a set of observations into a data set that is suitable for statistical analysis usually requires active data construction. If there are strong grounds to justify the necessary processing steps, the raw observations uniquely translate into a single data set for analysis. In many cases, however, the intermediate processing steps involve arbitrary or, as Leamer (1983) calls


Religiosity (Study 2)
R1
R2
R3
F1 F2 F3 F4 F5 F1 F2 F3 F4 F5 F1 F2 F3 F4 F5
0 0 0.03 0.02 0.08 0.02 0.02 0.03 0.05 0.06 0 0 0.02 0.01 0.04
0.01 0 0.08 0.03 0.16 0.07 0.05 0.18 0.24 0.41 0.01 0 0.09 0.06 0.23
0 0 0.06 0.04 0.37 0.02 0.03 0.07 0.08 0.21 0 0 0.05 0.03 0.23
0.01 0 0.13 0.08 0.44 0.06 0.03 0.22 0.24 0.52 0.01 0 0.14 0.09 0.43
0 0 0.03 0.01 0.08 0.15 0.07 0.17 0.07 0.14 0.02 0.01 0.06 0.02 0.07
0 0 0.02 0.01 0.06 0.2 0.05 0.42 0.23 0.44 0.03 0 0.14 0.05 0.19
0.01 0.01 0.05 0.01 0.1 0.39 0.2 0.45 0.11 0.26 0.08 0.04 0.17 0.03 0.13
0.01 0 0.05 0.02 0.11 0.33 0.09 0.59 0.26 0.55 0.09 0.02 0.26 0.08 0.27
0.01 0.01 0.02 0.09 0.31 0.11 0.09 0.38 0.26 0.78 0.02 0.02 0.11 0.12 0.5
0 0 0 0.06 0.07 0.07 0.1 0.08 0.14 0.19 0.01 0.02 0.02 0.06 0.08
0.02 0.01 0.06 0.11 0.36 0.06 0.04 0.3 0.13 0.66 0.02 0.01 0.13 0.07 0.46
0.02 0.01 0.02 0.15 0.13 0.04 0.05 0.07 0.07 0.16 0.01 0.02 0.03 0.05 0.09
0.05 0.03 0.12 0.07 0.19 0.17 0.11 0.47 0.32 0.7 0.06 0.04 0.23 0.12 0.38
0.01 0.01 0.01 0.05 0.03 0.06 0.11 0.05 0.17 0.14 0.02 0.03 0.02 0.07 0.04
EC1
EC2
ECL1
ECL2
ECL3
NMO1
NMO2
NMO3

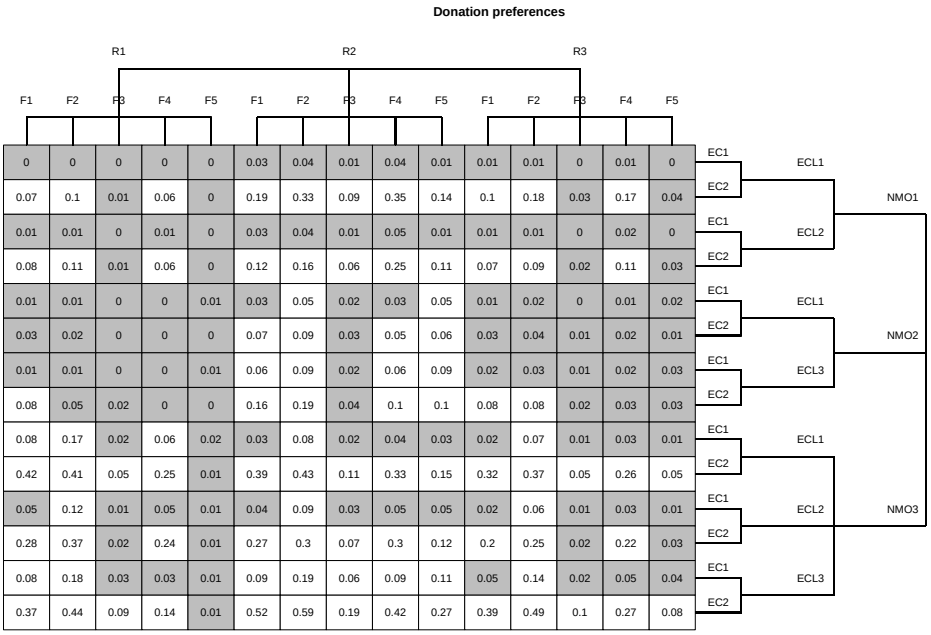
(a)


Social political attitudes
R1
R2
R3
F1 F2 F3 F4 F5 F1 F2 F3 F4 F5 F1 F2 F3 F4 F5
0 0 0 0 0 0.18 0.19 0.3 0.07 0.21 0.01 0.01 0.04 0 0.03
0.01 0.01 0.02 0 0.02 0.93 0.76 0.99 0.52 0.78 0.24 0.15 0.37 0.06 0.23
0 0 0 0 0 0.17 0.26 0.3 0.13 0.27 0.02 0.02 0.06 0.01 0.05
0.02 0.01 0.04 0 0.01 0.89 0.86 0.98 0.47 0.66 0.24 0.2 0.42 0.05 0.19
0.01 0.01 0.01 0 0 0.47 0.45 0.21 0.06 0.07 0.09 0.11 0.04 0.01 0.01
0.06 0.05 0.02 0.01 0 0.94 0.93 0.4 0.24 0.13 0.4 0.41 0.13 0.06 0.02
0.01 0.01 0 0 0 0.49 0.45 0.22 0.08 0.1 0.1 0.11 0.04 0.01 0.02
0.1 0.11 0.02 0.02 0 0.87 0.91 0.32 0.27 0.18 0.42 0.48 0.12 0.09 0.04
0 0 0.01 0 0.02 0.18 0.21 0.11 0.07 0.31 0.02 0.03 0.02 0.01 0.07
0.01 0.01 0.01 0 0.01 0.43 0.59 0.09 0.08 0.11 0.1 0.14 0.02 0.01 0.02
0.01 0 0.01 0 0.01 0.37 0.27 0.18 0.09 0.23 0.07 0.05 0.03 0.01 0.04
0.03 0.02 0.01 0.02 0.01 0.76 0.84 0.13 0.16 0.1 0.24 0.29 0.04 0.04 0.02
0.01 0.01 0.01 0 0 0.35 0.33 0.2 0.09 0.26 0.08 0.08 0.05 0.01 0.04
0.06 0.05 0.03 0.01 0 0.86 1 0.24 0.18 0.17 0.37 0.45 0.09 0.04 0.02
EC1
EC2
ECL1
ECL2
ECL3
NMO1
NMO2
NMO3

(b)



(c)



(d)

Figure 2.2: *Visualization of the multiverse of p-values of the fertility × relationship status interaction on religiosity (Panel A), on social political attitudes (Panel B), on voting preferences (Panel C) and on donation preferences (Panel D) in Study 2, showing the dependence of the results on data processing choices. See Box 1 for an explanation of the acronyms.*

them, whimsical, choices, so that the single set of observations does not uniquely lead to a single data set. Rather, it spawns a multiverse of data sets, and thus does not admit a unique conclusion. Yet, researchers often analyze, or at least report, only one (or a few) data sets that are the result of one (or a few) outcomes of this chain of arbitrary choices. To the extent their single data set is based on arbitrary processing choices, their statistical result is arbitrary. We suggest that, if several processing choices are defensible, researchers should perform a multiverse analysis instead of a single data set analysis. This involves considering *all* different reasonable data sets, except those arising under inconsistent choice combinations. A multiverse analysis is a way to avoid or at least reduce the problem of selective reporting, by making the fragility or robustness of the results transparent, and helps the identification of the most consequential choices.

In our demonstration, we started from a single set of raw data and performed both a single data set analysis as well as a multiverse analysis. Comparison of both types of analysis highlights the dramatic impact of going beyond an $N = 1$ sample from the multiverse. For religiosity in Study 1, the arbitrary data processing choices made in the single data set analysis led to a significant result. Placing this significant result in the multiverse of statistical results illustrates the risk of running a single data set analysis. The multiverse analysis revealed that almost all choice combinations for data processing lead to large p-values. As such non-significant findings in general represent nothing more than uncertainty, this pattern of results clearly raises serious questions regarding the finding on the effect of fertility found in the single data set analysis, and should make a researcher hesitant to trust the single data set finding. The effect of fertility on religion seems too sensitive to arbitrary choices and thus too fragile to be taken seriously.

For most other variables, there was considerable ambiguity: The interaction seemed to be significant across about half of the arbitrary choice

combinations. In these cases, the conclusion on the effect of fertility strongly depends on the evaluation of the different processing options. Both the authors performing the multiverse analysis, and the readers of the research can construct arguments in favor or against certain choices, and the validity of these arguments will help drawing the conclusion. For example, if additional information suggests that the fifth option of assessing fertility is clearly superior, then Panel A of Figure 2.2 indicates that there is little evidence for an effect of fertility on religiosity in Study 2. On the other hand, if additional information suggests that the second option of assessing fertility is clearly superior, then most choice combinations lead to a significant interaction effect.

If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large p-values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. One should reserve judgment and acknowledge that the data are not strong enough to draw a conclusion on the effect of fertility. The real conclusion of the multiverse analysis is that there is a gaping hole in theory or in measurement, and that researchers interested in studying the effect of fertility should work hard to *deflate* the multiverse. The multiverse analysis gives useful directions in this regard.

In general, deflating the multiverse involves developing a better and more complete theorizing of the constructs involved, and improving their measurement. Both routes for deflating the multiverse are illustrated in our case study. A first approach involves improving the experimental material and design. For example, the detailed multiverse examination shown in Figure 2.2 revealed that a lot of fluctuation hinged on the different choices for relationship status assessment. Thus, apparently, this type of research could benefit from a better way of assessing relationship status. Looking at the alternative options for assessing relationship status, it seems that

the ambiguous response Option 2 in the relationship status question could be formulated more precisely, so that relationship status assessment is no longer an arbitrary choice. This would have narrowed down the multiverses to 40 and 70 choice combinations in Study 1 and 2, respectively.

A second approach for deflating the multiverse involves developing more complete and more precise theory, in such a way that some options are theoretically superior than others, and it should be preferred when constructing data sets. For example, a great deal of variation in the results appeared to be driven by the different options for assessing fertility. Clearly, for this type of research, developing and applying a more precise way of assessing fertility should become a research priority. The availability of different reasonable options for estimating next menstrual onset or for classifying women into a high or low fertility group, based on their cycle day, stems from the fact that a precise theoretical foundation is lacking (Harris, 2013). The development of elaborated theories concerning these issues would narrow down the number of alternative options and deflate fluctuation. Recently, Gangestad et al. (2016) have recommended assessing fertility based on the detection of surges in luteinizing hormone, ideally in a within-subjects design. It is of note that this alternative strategy of assessing fertility was used in several papers by Durante (e.g., Durante et al., 2011, 2012).

Preregistration (e.g., Chambers, 2013; Wagenmakers et al., 2012) or blind analysis (e.g., MacCoun & Perlmutter, 2015) are not useful strategies for deflating the multiverse. By preregistering a study, all analytical choices—including the arbitrary ones—are made ahead of time, before collecting the data. Similarly, in a blind analysis, all analytical choices are made using a data set with temporarily removed data labels. The appeal of both strategies is that the choices cannot be made conditional on the (real) data. However, the considered results are still just the results given one choice combination, albeit preregistered or blindly made, and their ro-

bustness across other reasonable choice alternatives remains hidden from view. Thus, pre-registration or blind analysis do not preclude a multiverse analysis, as they do not annihilate the arbitrariness in data preparation.

As is evident from our demonstration, a multiverse analysis is highly context-specific and inherently subjective. Listing the alternative options for data construction requires judgement about which options can be considered reasonable and will typically depend on the experimental design, the research question and the researchers performing the research. Whereas this subjectivity may seem undesirable, presenting results given only a single combination of reasonable options is much more misleading; indeed one of the sources of the current crisis in scientific replication is that researchers traditionally have taken p-values at face value without considering the multiplicity of choices in data construction.

A related point is that not all options are necessarily exactly interchangeable. Some options might seem better than others, at least for some researchers. If such is the case, this knowledge can be used to construct arguments for interpreting results such as those shown in Figure 2.2. However, a multiverse analysis should involve all plausible construction alternatives, not just the most plausible ones. Only when one choice is clearly and unambiguously the most appropriate one, variation across this choice is uninformative.

The richness of possibilities for different data processing choices present in the raw data made the case study exceptionally suitable for the demonstration of a multiverse analysis. We do not expect that all multiverses will consist of such a numerous amount of data sets. The fact that more typical multiverses will tend to be smaller does not make a multiverse analysis less necessary. Even when confronted with only one arbitrary data processing choice, researchers should be transparent about it and reveal the sensitivity of the result to this choice.

We aimed to show the multiverse analysis we think Durante et al. (2013) could have done, instead of their single data set analysis. As their single data set analysis used p-values, our demonstration of the multiverse analysis did too. There is, however, nothing inherently special about p-values from a multiverse perspective. Increasing the transparency in reporting through a multiverse analysis is valuable, regardless of the inferential framework (frequentist or Bayesian), and regardless of the specific way uncertainty is quantified: a p-value, an effect size, a confidence (Cumming, 2013) or credibility (Kruschke, 2010) interval or a Bayes factor (Morey & Rouder, 2011).

The primary goal of a multiverse analysis is to enhance research transparency. Unlike, for example, a p-curve analysis (Simonsohn, Nelson, & Simmons, 2014), it is not a formal test of questionable research practices such as selective reporting, or a method to estimate the strength of the evidence for an effect. The multiverse analysis does not produce a single value summarizing the evidential value of the data, nor does it imply a threshold for an effect to reach to be declared robustly significant. Nevertheless, one might try to summarize the multiverse analysis more formally. One reasonable first step is to simply average the p-values in the multiverse, in this case averaging all the numbers displayed in Figure 2.1 or 2.2. This mean value can be considered as the p-value of a hypothetical preregistered study with conditions chosen at random among the possibilities in the multiverse and seems like a fair measurement in a setting where all of the possible data-processing choices seem plausible (as in the example presented here, where the different options are drawn from other papers in the relevant literature).

We have focused on the multiverse of statistical results originating from the data multiverse (i.e., the different reasonable choices in data processing). We have ignored arbitrary choices occurring at the level of statistical models used in data analysis. Choices at the model level include choosing among

different statistical approaches (e.g., a repeated-measures ANOVA or a hierarchical linear model), focusing on main effects or interactions, approximating errors normally, assuming random effects, assuming homoscedasticity, assuming linearity, choosing between a parametric and a non-parametric approach, and so on. One specific analysis thus corresponds to a single sample from a *model multiverse*. If the choice for a single model specification out of the model multiverse cannot be justified, a model multiverse analysis can be performed to reveal the effect of this arbitrary choice on the statistical result.

A compelling example of such a model multiverse is provided in Patel, Burford, and Ioannidis (2015), focusing on the choices in deciding which predictors and covariates to include. Such a model multiverse analysis is related to *perturbation analysis* (Geisser, 1993) and to *sensitivity analysis* in economics (e.g., Leamer, 1985) and in Bayesian statistics (e.g., Kass & Raftery, 1995), all of which involve investigating the influence of arbitrary modeling assumptions on the results, such as using a normal error distribution or a t-distribution, the inclusion of different variables, or using different reasonable priors. In a more complete analysis, the multiverse of data sets could be crossed with the multiverse of models to further reveal the multiverse of statistical results. Thus, the multiverse analysis as demonstrated here is a minimal attempt at establishing a range of analyses consistent with a research hypothesis. To the extent that there are arbitrary choices not only in data preparation but also in data analysis or model choice, this motivates encompassing analyses of multiple predictors, interactions, or outcomes in a hierarchical model so as to reduce problems of multiple comparisons (Gelman, Hill, & Yajima, 2012).

Our demonstration of the multiverse analysis should serve as a cautionary tale. We hope it raises awareness that, in the light of the multiverse of statistical results, isolating a single statistical result stemming from a chain

of arbitrary choices can be highly misleading. Readers of research need to get a sense of the sensitivity of conclusions to arbitrary decisions in data preparation, and thus of the fragility or robustness of a claimed effect. We believe that it should become standard practice to go beyond a single data set analysis and to acknowledge the multiverse of statistical results. Admittedly, performing a multiverse analysis will often be difficult, and to a large extent subjective, but that does not change the fact that it is a necessary step for increasing transparency.

Appendix

For the demonstration of the multiverse analysis in this chapter, we focused on Study 1 and Study 2 from Durante et al. (2013). Kristina Durante kindly provided the raw data, which were collected with a survey, the processed data, and the research materials (i.e., the survey questions), and gave us the permission to make them publicly available. The data (and the code) can be found on `https://osf.io/zj68b/`. In the following, we will give a description of these files.

Our single data set analysis used the same processing choices as Durante et al. (2013). Note that Durante et al. (2013) write that women in both studies had regular monthly menstrual cycles (25–35 days), suggesting that they excluded women with cycle lengths shorter than 25 or longer than 35 days. However, looking at the data file, it seems that they did not exclude participants based on this criterion. When we do not exclude these women, our single data set analysis arrives at the exact same results as theirs.

Raw Data

Study 1

- WorkerID. ID of participant
- Answers to religiosity items. For items 2 and 3: "Please indicate how much you agree with the following statement".
  - Rel1: "How much do you believe in God?" 1– 9
  - Rel2: "I see myself as a religiously oriented person." 1 – 9
  - Rel3: "I believe that God or a Higher Power is responsible for my existence." 1 – 9
- Date Testing. Date of participant filling in the questionnaire.

- Answers to questions about menstrual period. "Please use the calendars to answer the following questions." (The calendars are not reproduced here.)
  - Start Date of Last Period: "Please give your best estimate of the date on which you started your last period (please be as precise as possible). This date was probably within the last few weeks. Sometimes thinking of where you were when you started your last period helps. For instance, was it on a weekend?, were you at work, was it during a football game?, etc. Please write the date in mm/dd/yyyy format (e.g., 8/18/2012)."
  - Sure 1: "How sure are you about that date?" 1 – 9. This variable was included in the data file, but it was not used in Durante et al.'s (2013) analyses.
  - Start Date of Period Before Last: "Please give your best estimate of the date on which you started the period before your last period (please be as precise as possible). Please write the date in mm/dd/yyyy format (e.g., 7/18/2012)."
  - Sure 2: "How sure are you about that date?" 1 – 9. This variable was included in the data file, but it was not used in Durante et al.'s (2013) analyses.
  - Cycle Length: "How many days long are your menstrual cycles? (for most women, the range is between 25-35 days) Keep in mind this is the number of days from the start of one menstrual period to the start of the next menstrual period and NOT the length of your menstrual bleeding." This variable was included in the data file, but it was not used in Durante et al.'s (2013) analyses.
- Relationship. "What is your current romantic relationship status?" (1) *not dating/romantically involved with anyone*, (2) *dating or in-*

volved with only one partner, (3) *engaged or living with my partner*, (4) *married*, or (5) *other*. If participants picked response (5), they were prompted to provide a description of their relationship, which was subsequently coded into one of the four options by the original authors. In the data made available to us, all (5) responses were already coded into another response option. In this sense, the data we start from in our multiverse analysis do not fully correspond to the raw data. Further, Durante et al. (2013) describe response Option (2) as *dating*, but the materials indicate *dating or involved with only one partner* was used.

- The following additional raw variables were included in the survey, but were not used in the analysis reported by Durante et al. (2013) or in this chapter: Age ("How old are you?"); Ethnicity ("What is your ethnicity?"); Income ("What is your current household income?"); Children ("Do you have children?"); "What Country/State do you live in?". The responses to the last question were not included in the data file we received.

Study 2

The raw data included the same variables as in Study 1, plus the following variables.

- Answers to fiscal political attitudes items ("Please indicate how much you agree with the following statements.")
  - RichTax: "The rich should pay a higher tax rate than the middle class." 1 – 7
  - TooMuchProfit: "Business corporations make too much profit." 1 – 7

  - StandardLiving: "Government should ensure that all citizens meet a certain minimum standard of living." 1 – 7
  - FreeMarket: "In nearly every instance, the free market allocates resources most efficiently." 1 – 7
  - PrivSocialSec: "Privatize Social Security." 1 – 7

- Answers to social political attitudes items ("Please indicate how much you agree with the following statements.")
  - Abortion: "Abortion is a women's [sic] right." 1 – 7
  - Marriage: "Marriage is between a man and a woman." 1 – 7
  - StemCell: "Stem cell research is moral and can be useful for science." 1 – 7
  - Marijuana: "Marijuana should be legal." 1 – 7
  - RestrictAbortion: "Laws should restrict abortion in all or most cases." 1 – 7

- Vote. "Imagine walking into the voting booth today. Who would you vote for in the presidential election?" Mitt Romney (republican) – Barack Obama (democrat)

- Donate. "For the next part of the study we will donate $1 to the presidential campaign of your preferred candidate. Please indicate which candidate's campaign you would like us to donate $1 to." Mitt Romney — Barack Obama

- Start Date Next. The research material we received does not contain a question about the variable. Durante et al. (2013) write that they asked participants to indicate the expected start date of their next menstrual period in both studies, but only the data file for Study 2 contained this variable. This variable was included in the data file, but it was not used in Durante et al.'s (2013) analyses.

Processed Data

Study 1

The data file we received contained the following processed variables.

- Religiosity Score. Average of Rel1, Rel2 and Rel3
- Cycle Day. Variable indicating each participant's estimated cycle day, ranging from 1 to 28. There was no documentation on how this variable was calculated, but we managed to reconstruct this variable using the following rules:
  - Cycle Day = 28 minus Days Before Next Menstrual Onset (Cycle Day $< 1 = 1$, Cycle Day $> 28 = 28$)
  - Days Before Next Menstrual Onset = Next Menstrual Onset minus Date Testing
  - Next Menstrual Onset = Start Date of Last Period plus Computed Cycle Length
  - Computed Cycle Length = Start Date of Last Period minus Start Date of Period Before Last
- Fertility. Variable indicating each participant's fertility (high or low) based on Cycle Day. High = Cycle Day 7 – 14; Low = Cycle Day 17 – 25.
- Relationship Status. Variable indicating each participant's relationship status based on the raw variable Relationship. Single = response Option (1) or (2); Married = response Option (3) or (4).
- Cycle Day Testing. There was no documentation on how this was calculated, but we managed to reconstruct it as Date Testing minus Start Date of Last Period plus 1. This variable was not directly used

in Durante et al.'s (2013) analyses. It might have been used as an intermediate step in the calculation of Cycle Day using other rules than the ones described above.

Study 2

The processed variables in the data file were the same as in Study 1, plus the following.

- Fiscal Political Attitudes Score: Average of FreeMarket, PrivSocialSec, RichTax, StandaardLiving and TooMuchProfit.
  Note: For some of the fiscal political attitudes items (e.g., Profit), answers were first reversed, such that higher values indicated conservatism and lower values indicated liberalism.

- Social Political Attitudes Score: Average of Marriage, RestrictAbortion, Abortion, StemCell, and Marijuana.
  Note: For some of the social political attitudes items (e.g., Abortion), answers were first reversed, such that higher values indicated conservatism and lower values indicated liberalism.

- ConceptionProbability: Variable indicating each participant's conception probability based on the variable Cycle Day. This variable was not used in the analyses that we focus on in this chapter.

- Cycle Day: For all but two participants, we managed to reconstruct this variable using the same rules as in Study 1 (see above), after fixing some coding errors (see below).

Data Cleaning

Some raw variables contained obvious coding errors. For some values, it was possible to fix the errors, whereas for others, we could not provide fixes with reasonable confidence. In these cases, erroneous values were set to NA (not available).

- Some values of the raw variables Start Date of Last Period, Start Date of Period Before Last, and Start Date Next indicated the wrong year (i.e., something else than 2012, the year in which the questionnaire was filled out, such as 2010, 2011, 2013, 2022, 2912). In these cases, we fixed the year to 2012.

- For one participant, the Start Date of Last Period and the Start Date of Period Before Last were identical. In this case, we used the value of Cycle Day to fix the Start Date of Period Before Last.

- Some values of Start Date Next were before the date of testing, and so were obviously wrong. They were converted to NA.

- Some values of Cycle Length were unusually small (e.g., 4, 5, 6, or 7) or unusually large (e.g., 90, 40972, or 41035). They were converted to NA.

- Some values of Cycle Length were expressed using a range (e.g., 21-26, about 31 days, 40+). They were converted to NA.

- For two participants, we did not manage to recover the value of Cycle Day from the original data file. As this likely indicates a coding error in Start Date of Last Period or Start Date of Period Before Last, these variables were converted to NA. This means that for these two participants, Next Menstrual Onset, and hence Cycle Day and Fertility, could only be determined under NMO3 from Box 1 in this chapter.

However, when Cycle Day was determined based on NMO1 from Box 1 in this chapter, we used the processed variable Cycle Day from the original data file for the assessment of fertility to ensure that the results of our single data set analysis are identical to the single data set analysis in Durante et al. (2013), despite these coding errors.

Chapter 3

A theoretical note on the prior information criterion

Introduction

In psychology, as well as in other scientific fields, statistical models are used to describe the underlying processes of probabilistic phenomena under study and thereby explain the regularities behind observed data. In developing such models, it is often the case that different plausible accounts are proposed. Toward providing an objective criterion for testing competing models, the development of model selection methods has been an important topic of research (e.g., Claeskens & Hjort, 2008). With this goal in mind, van de Schoot et al. (2012) proposed the *prior information criterion* (PIC) as a Bayesian method for testing models under (in)equality constraints (e.g., whether a certain parameter is greater than, less than, or equal to a fixed value).

For a model with parameter θ, likelihood $f(y|\theta)$ and prior $p(\theta)$, the PIC is defined as

$$\begin{aligned} \text{PIC} &= E_{p(\theta)}\left[-2\log f(y|\theta)\right] \\ &= -2\int \log\left(f(y|\theta)\right) p(\theta)\,\mathrm{d}\theta. \end{aligned} \tag{3.1}$$

As $\log f(y|\theta)$ is often used to indicate model-data fit, the PIC reflects a lack

of fit. Selection between two models is based on the difference in their PIC values, $\Delta\text{PIC} = \text{PIC}_1 - \text{PIC}_2$:

$$\Delta\text{PIC} = -2 \int \log\left(f_1(y|\theta_1)\right) p_1(\theta_1)\,\mathrm{d}\theta_1 + 2 \int \log\left(f_2(y|\theta_2)\right) p_2(\theta_2)\,\mathrm{d}\theta_2. \tag{3.2}$$

A negative value of ΔPIC indicates a preference for Model 1 over Model 2, whereas a positive value indicates a preference for Model 2 over Model 1.

As noted by van de Schoot et al. (2012), ΔPIC is closely related to the Bayes factor (BF; Jeffreys, 1961). The Bayes factor is the ratio of two marginal likelihoods (ML), with the ML defined as

$$\begin{aligned} \text{ML} &= E_{p(\theta)}\left[f(y|\theta)\right] \\ &= \int f(y|\theta) p(\theta)\,\mathrm{d}\theta. \end{aligned} \tag{3.3}$$

The marginal likelihood integrates over the entire prior distribution, and thus reflects the average fit of a model to the data, weighted by the prior beliefs about the model parameter θ. In this chapter, we will express the Bayes factor $\text{BF} = \frac{\text{ML}_1}{\text{ML}_2}$ in the following form: $-2\log\text{BF} = -2\log\text{ML}_1 + 2\log\text{ML}_2$. This way, we obtain a model selection criterion on the same scale as the well-known likelihood-ratio test statistic or the deviance (Kass & Raftery, 1995):

$$-2\log\text{BF} = -2\log\int f_1(y|\theta_1) p_1(\theta_1)\,\mathrm{d}\theta_1 + 2\log\int f_2(y|\theta_2) p_2(\theta_2)\,\mathrm{d}\theta_2. \tag{3.4}$$

As with ΔPIC, a negative value of $-2\log\text{BF}$ indicates that Model 1 is more likely than Model 2, whereas a positive value indicates the opposite.

A clear commonality between the PIC and the ML is their sensitivity to the prior. While this sensitivity is sometimes seen as a drawback, it can also be considered an advantage. One major form of Bayesian hypothesis testing involves specifying different priors for a given likelihood, each corresponding to different hypotheses, and comparing the resulting models given observed data. In this case, sensitivity to the prior is an advantage rather than a

drawback (e.g., Vanpaemel, 2010). For example, using the PIC or the Bayes factor, one can check whether a regression coefficient is greater than a certain value by testing a model that incorporates such a hypothesis in the form of a constrained prior.

Besides their similar forms, ΔPIC and $-2\log\text{BF}$ are two distinct model selection criteria. Their technical difference is apparent from the location of the logarithm operator: The PIC places the logarithm inside the integral (Equation 3.2) whereas the BF places it outside the integral (Equation 3.4). In this chapter, we illustrate applications in which this difference leads to very different model selection outcomes of the two criteria. In particular, we report that the application of the PIC produces unexpected and puzzling results, which can be explained based on its analytic forms. Further, we present a formal relationship between the two criteria for general cases and provide insight into their disparate behavior.

Application to Binomial Models

We investigate the behavior of the PIC and the Bayes factor in the context of the very basic, but widely used, binomial model with a single parameter θ representing the probability of success. In creating multiple scenarios of hypothesis testing, both inequality and equality constraints on the parameter are considered. Although the original motivation behind the PIC's development was to propose a method for testing inequality constrained hypotheses, van de Schoot et al. (2012) themselves applied the PIC to test an equality constrained hypothesis as well (e.g., Real-Life Example 1, pp. 14–15, van de Schoot et al., 2012). Moreover, we see no rationale in the derivation of the PIC that fundamentally precludes its application in testing equality constraints.

Models and Analytic Expressions

Consider the number of successes $y = 0, 1, \ldots, n$ with the binomial probability mass function,

$$f(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}, \tag{3.5}$$

where $0 < \theta < 1$ is the probability of success. We focus on four models representing different hypotheses about θ, which are specified by different priors for θ: an unconstrained model ($M_{01} : 0 < \theta < 1$), a one-sided inequality constrained model ($M_{0u} : 0 < \theta < u$), a two-sided inequality constrained model ($M_{lu} : l < \theta < u$), and an equality constrained model ($M_c : \theta = c$). The unconstrained and inequality constrained models assume uniform distributions whose supports are defined by their hypothesized intervals. The equality constrained model places a point mass at c (see Figure 3.1).

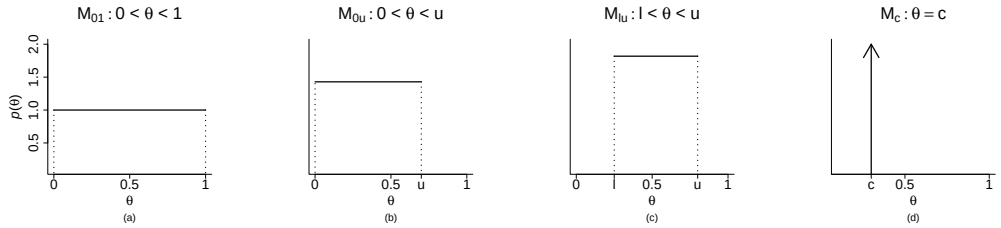


Figure 3.1: *Prior distributions for θ, specifying the unconstrained model (a), one-sided inequality constrained model (b) and two-sided inequality constrained model (c) and equality constrained model (d).*

Given these models, our analysis of the behavior of the PIC and the Bayes factor considers three scenarios in which each of the constrained models is tested against the unconstrained model (i.e., M_{0u} vs. M_{01}, M_{lu} vs. M_{01}, and M_c vs. M_{01}). To conduct these tests, we derived analytic-form expressions of the PIC and the ML for each model (see Appendix). These are summarized in Table 3.1 and illustrated in Figure 3.2 in the forms of ΔPIC and $-2 \log \text{BF}$ for each scenario. In particular, Figure 3.2 displays the behavior of ΔPIC

and $-2 \log \text{BF}$ applied to the test of the constrained models against the unconstrained model for every possible sample proportion y/n when $n = 100$. For each type of the constrained hypotheses (each row), three examples of constraints (each column) are considered. For each method, a positive value indicates a preference for the constrained model over the unconstrained model, and a negative value does the opposite, as denoted on the leftmost vertical axis.

Testing One-Sided Inequality Constrained Hypotheses

The top row of Figure 3.2 shows the values of ΔPIC and $-2 \log \text{BF}$ as a function of observed proportions when testing a one-sided inequality constrained hypothesis ($M_{0u} : 0 < \theta < u$) against the unconstrained model ($M_{01} : 0 < \theta < 1$). As seen in the figure, in this comparison scenario, both criteria decrease with the increasing sample proportion y/n, which makes good sense. Both methods select M_{01} if high values of y/n are observed and select M_{0u} if low values of y/n are observed. Moreover, consistent with what would be expected, both methods prefer M_{01} more often if M_{0u} places a more strict constraint on θ (i.e., a smaller value of u).

Despite these similarities, there are two clear differences between the two methods. First, the trend of ΔPIC over y/n is linear whereas that of $-2 \log \text{BF}$ is curved. The second difference concerns the point on y/n where their values cross zero, or the *decision bound*, at which neither of the two models is clearly supported by the data. Consider the case of ΔPIC first. Analytic solutions for ΔPIC's decision bounds for each model comparison are listed in Table 3.2. Inspection of the table reveals that the decision bound of ΔPIC on y/n is constant and not affected by the sample size n. This fact provides a perspective from which to better view the behavior of ΔPIC seen in Figure 2. When M_{0u} with $u = .3$ is compared against M_{01} (upper left panel in Figure 3.2), the decision bound of ΔPIC is approximately

Table 3.1: *Analytic expressions of* ΔPIC *and* $-2\log\text{BF}$ *applied to three pairwise comparisons of four binomial models with different priors: The unconstrained model (*M_{01}*) with a uniform prior on the interval* $[0, 1]$*, the one-sided inequality constrained model (*M_{0u}*) with a uniform prior on the interval* $[0, u]$*, the two-sided inequality constrained model (*M_{lu}*) with a uniform prior on the interval* $[l, u]$*, and the equality constrained model (*M_c*) specified by a Dirac delta function at* c*. The derivations can be found in the Appendix.*

M2 vs. M1	ΔPIC
$M_{0u} : 0 < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$2y\big(\log(u) - \frac{u-1}{u}\log(1-u)\big) + 2n\big(\frac{u-1}{u}\log(1-u)\big)$
$M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$-2y\frac{-(1-l)\log(1-l)+u\log(u)+(1-u)\log(1-u)-l\log(l)}{u-l} - 2n\frac{(1-l)(\log(1-l)-1)-(1-u)(\log(1-u)-1)}{u-l}$
$M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$	$2y\left(\log\frac{c}{1-c}\right) + 2n\big(1 + \log(1-c)\big)$
M2 vs. M1	$-2\log\text{BF}$
$M_{0u} : 0 < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$2\log\frac{I_u(y+1,n-y+1)}{u}$
$M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$2\log\frac{I_u(y+1,n-y+1)-I_l(y+1,n-y+1)}{u-l}$
$M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$	$2y\left(\log\frac{c}{1-c}\right) - 2\log B(y+1, n-y+1) + 2n\big(\log(1-c)\big)$

Note: y is the observed number of successes, n is the number of trials, l, u and c are constants, $B(\cdot,\cdot)$ is the beta function, $B_u(\cdot,\cdot)$ is the incomplete beta function, and $I_u(\cdot,\cdot)$ is the regularized incomplete beta function.

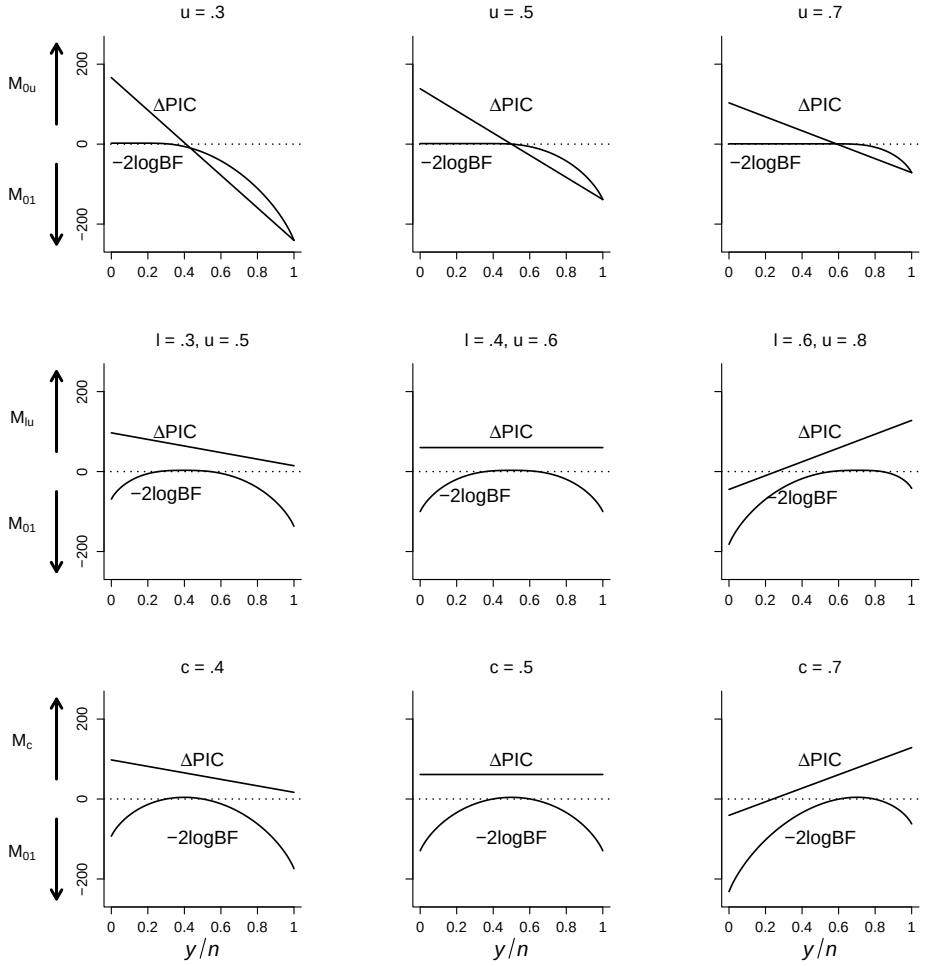


Figure 3.2: *Pairwise comparison of binomial models under different hypotheses based on the PIC and BF when* $n = 100$*. For both* ΔPIC *and* $-2\log\text{BF}$*, a positive value indicates a preference for the constrained model and a negative value indicates a preference for the unconstrained model.*

Top row: Test of the one-sided inequality constrained model ($M_{0u} : 0 < \theta < u$) *against the unconstrained model* ($M_{01} : 0 < \theta < 1$)*.*

Middle row: Test of the two-sided inequality constrained model ($M_{lu} : l < \theta < u$) *against the unconstrained model* ($M_{01} : 0 < \theta < 1$)*.*

Bottom row: Test of the equality constrained model ($M_c : \theta = c$) *against the unconstrained model* ($M_{01} : 0 < \theta < 1$)*.*

Table 3.2: *Decision bounds of* ΔPIC *for the three pairwise comparisons. At decision bounds, the analytic expressions of* ΔPIC *in Table 3.1 equal zero, indicating no model preference.*

M2 vs. M1	Decision bound
$M_{0u} : 0 < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$-\frac{-(1-u)\log(1-u)}{(1-u)\log(1-u)+u\log(u)}$
$M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$	$-\frac{(1-l)\log(1-l)-(1-u)\log(1-u)}{-(1-l)\log(1-l)+(1-u)\log(1-u)+u\log(u)-l\log(l)}$
$M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$	$-\frac{1+\log(1-c)}{\log\frac{c}{1-c}}$

.41, meaning that when the observed proportion y/n is less than .41, the constrained model is favored. According to the closed-form expressions in Table 3.2 (though not visible in Figure 3.2), this will be the case no matter what the sample size n is.

The fact that the decision bound of ΔPIC is insensitive to n has a significant implication for its behavior that is distinguished from that of $-2\log \mathrm{BF}$. Suppose that the underlying process is a binomial model whose success probability θ_0 lies above the upper bound u of the constraint, but below the decision bound (e.g., $\theta_0 = .35$ when $u = .3$). Then, it is reasonable to expect that, once a sufficient amount of data is collected, the unconstrained model ($M_{01} : 0 < \theta < 1$) should almost surely be identified against the incorrect constrained model ($M_{0u} : 0 < \theta < u$). The selection outcome of ΔPIC contradicts this intuition: As n increases, by the central limit theorem, the observed proportion y/n concentrates at $\theta_0 = .35$ and since this is below the decision bound (i.e., .41), ΔPIC will incorrectly select the constrained model. Thus, even when the truth lies *only* in one model and not in the alternative one, the PIC can prevent one from selecting the model containing

the truth.[1]

By contrast, in the same situation, $-2\log \text{BF}$ behaves as anticipated. This can be seen as follows. In the expression of $-2\log \text{BF}$ for testing M_{0u} versus M_{01}, shown in Table 3.1, the numerator of the fraction inside the logarithm (i.e., $I_u(y+1, n-y+1)$) is in fact $p(\theta < u|y)$, the posterior probability of θ being less than u. Since, as per the large-sample property of Bayesian posteriors (Schervish, 1995), the posterior distribution of θ concentrates at the true proportion θ_0 as n increases, $p(\theta < u|y)$ converges to 1 when $\theta_0 < u$ and to zero when $\theta_0 > u$. Therefore, the selection based on $-2\log \text{BF}$ will be consistent with the underlying process as data accumulate.

Testing Two-Sided Inequality Constrained Hypotheses

The test of a two-sided inequality constrained hypothesis (i.e., $M_{lu} : l < \theta < u$ vs. $M_{01} : 0 < \theta < 1$) is shown in the middle row of Figure 3.2. In this comparison scenario, the PIC and the BF exhibit more drastic differences in their selection behavior. Again, ΔPIC's linear versus $-2\log \text{BF}$'s nonlinear relationships to the data y/n are apparent. In this case, however, a selection criterion's nonlinear response to the evidence in data is vital for its outcome to make sense: It is supposed to select M_{01} when the observed proportion

[1] One may wonder if the behavior of ΔPIC may be due to a particular specification of the prior distribution (i.e., a uniform distribution in the current example). The decision bound indeed depends on the prior. In fact, in the case of $u = .5$ under the uniform prior, the decision bound happens to be .5, which will prevent counterintuitive selection behavior. Even for another value of u, the decision bound can be made to avoid such illogical selection by using a different prior (our derivation of the PIC in the Appendix assumes a general class of beta priors). However, a proper Bayesian inference should not require one to confine models to a particular prior specification in order to achieve consistency in model selection. Normally, the effect of priors is overridden by a sufficient accumulation of evidence in data. This is precisely the behavior of Bayes factors, which has been proved for general cases (Doob, 1949; Schwartz, 1965), but, as shown in our analysis, is not exhibited by ΔPIC.

is sufficiently far away from the hypothesized range of M_{lu} ($l < \theta < u$) by being either small or large, and it should favor M_{lu} when the observation is within or near the range.

However, ΔPIC's selection behavior belies this intuition. When testing $M_{lu} : .3 < \theta < .5$ (leftmost panel), ΔPIC always favors M_{lu} no matter what proportion is observed, even when it is very small or very large. In the case of testing $M_{lu} : .4 < \theta < .6$ (middle panel), the quantity itself of ΔPIC is completely insensitive to the data as it is reduced to a positive constant. The closed-form expression in Table 3.1 shows that this will always be the case whenever $u = 1 - l$. When $M_{lu} : .6 < \theta < .8$ is tested (rightmost panel), in which the hypothesized range shifts closer to an extreme proportion, ΔPIC finally allows for the possibility that M_{01} is selected, but only when very small proportions are observed. Still, ΔPIC does not accept any evidence for M_{01} when the proportion lies beyond the other side of the constraint. For instance, ΔPIC prefers $M_{lu} : .6 < \theta < .8$ even with 100 successes out of 100 trials.

Note that, unlike in the previous comparison scenario involving one-sided constraints, we do not need to postulate a large-sample condition in order to see the PIC's counterintuitive selection behavior. For the test of two-sided inequality constraints, its illogical selection can occur for *any* data, regardless of the sample size. To see this, suppose that the data-generating process is a binomial model with θ_0 located outside the interval (l, u) so that we expect that with sufficient data M_{01} should almost surely be preferred over M_{lu}. For example, suppose that the underlying truth is $\theta_0 = .2$ or $\theta_0 = .7$ when $M_{lu} : .3 < \theta < .5$ is tested (e.g., consider the first example in the middle row of Figure 2). In this situation, ΔPIC, as a linear function of y, has a decision bound beyond 1, which is the logical limit of y/n. Since Table 3.2 shows the decision bound is not affected by n, the PIC will never select M_{01}, no matter what data in any size sensibly support M_{01}.

More generally, it is the very nature of the constraint $l < \theta < u$, bounded on two sides, that prevents the PIC from handling such a hypothesis properly. Whenever an inequality constraint of this type is tested, simply because ΔPIC, as a linear function of data, cannot form two decision bounds, there always exists an underlying proportion θ_0 outside the interval (l, u) that makes it impossible for the PIC to favor M_{01} with any data of any amount. In some cases, the PIC will never select M_{01} for all values of the underlying truth θ_0 outside (l, u) no matter what data evidence M_{01} (e.g., examples in the leftmost and middle panels of the figure), and in other cases, data in support of M_{01} cannot arise if θ_0 is either very small or very large (e.g., the example in the rightmost panel).

In contrast, again, the BF performs as expected: $-2 \log \text{BF}$ responds to the data nonlinearly, favoring M_{01} when the sample proportion is far from the two-sided constraint (l, u), and M_{lu} when the observation is in or close to (l, u). This is reflected in its inverted U-shapes in all three examples of (l, u) in Figure 3.2. Asymptotically, in the expression of $-2 \log \text{BF}$ for testing M_{lu} versus M_{01} in Table 3.1, the numerator inside the logarithm equals $p(l < \theta < u|y)$, which converges to 1 when $l < \theta_0 < u$ and to zero when $\theta_0 < l$ or $\theta_0 > u$ as n increases. Therefore, the decision based on $-2 \log \text{BF}$ is consistent.

Testing Equality Constrained Hypotheses

The results from testing an equality constrained hypothesis (i.e., $M_c : \theta = c$ vs. $M_{01} : 0 < \theta < 1$) are shown in the bottom row of Figure 3.2. The selection pattern of the two methods across the data y/n is qualitatively the same as in the previous scenario of testing two-sided inequality constrained hypotheses illustrated in the middle row of the figure. With the hypothesis $\theta = .4$ (leftmost panel), ΔPIC is above zero irrespective of data, thus always selecting M_c. When testing $\theta = .5$ (middle panel), ΔPIC becomes a positive

constant, again always selecting M_c. Only when the hypothesized value for θ is away from .5 to some extent like $c = .7$ (rightmost panel), ΔPIC allows M_{01} to be selected for very small y/n (decision bound $\approx .24$), but not for large proportions.

Essentially, the properties of the two methods described previously still hold here: A situation exists in which the PIC will never favor M_{01}, no matter what data in any size sensibly support M_{01}, whereas the BF performs as expected. In testing the equality constraint $\theta = c$, a selection criterion is expected to prefer M_{01} when y/n is away from the hypothesized value c by being either small or large, and it should favor M_c when y/n is equal or close to c. ΔPIC cannot accomplish this by nature, because it is a linear function of data and cannot provide two decision bounds. In addition, as with earlier tests, the decision bound of ΔPIC for M_c versus M_{01} is constant for all sample sizes as shown in Table 3.2. Consequently, there always exists an underlying proportion $\theta_0 \neq c$ that makes it impossible for the PIC to select M_{01} given any data. In some cases (e.g., $c = .4$ or $c = .5$; i.e., leftmost and middle panels), the PIC will never choose M_{01} for all values of the underlying truth $\theta_0 \neq c$ no matter what data support M_{01}. In other cases, data evidencing M_{01} cannot arise if θ_0 is either very small or very large (e.g., in the rightmost panel).

The BF performs as a reasonable criterion should, selecting M_c only when the observed proportion is close to c, the hypothesized value of θ. Its selection is again consistent, which can be seen in the analytic form of $-2 \log \text{BF}$ for M_c versus M_{01} in Table 3.1. In this case, $-2 \log \text{BF}$ equals $2 \log p(\theta = c|y)$, twice the log posterior density function of θ evaluated at c, which goes to ∞ if $\theta_0 = c$, and to $-\infty$ if $\theta_0 \neq c$ as n increases, due to the asymptotic convergence of a posterior distribution.

Generalization

We have shown two peculiar properties of the PIC when applied to the comparison of a binomial model with an inequality or an equality constraint on its success probability against the encompassing, unconstrained model. First, ΔPIC is always a linear function of the observed proportion, and as a result, provides one-sided evidence, responding to only either small or large proportions in support of one of the models. Second, ΔPIC is insensitive to the accumulation of data in the sense that its decision bound remains constant for all sizes of a binomial sample. As a consequence, for a certain range of observed proportions, the PIC cannot favor one of the two models even when the data reasonably support it, no matter how large a sample is collected. In a subset of the cases, it can be shown that the unconstrained model is never selected with any supportive data of any size since ΔPIC is signed towards the constrained model over all possible sample proportions. By contrast, in all of these testing scenarios, the Bayes factors perform as expected, selecting the model that receives sensible support from the data. The selection of Bayes factors is also consistent, recovering the underlying process as the sample size increases.

In this section, we examine distinctive properties of the PIC in more general cases. It turns out that the same behavior that hallmarks the PIC when applied to binomial models holds for the general exponential family distributions, which include many standard probability distributions in statistics, such as Bernoulli, binomial, Poisson, normal, and so forth (Schervish, 1995). When the PIC is applied to Models 1 and 2 constructed by imposing two different priors on an exponential family, $\Delta\text{PIC} = \text{PIC}_1 - \text{PIC}_2$ has the form

$$\begin{aligned}\Delta\text{PIC} = 2\left(\sum_{i=1}^{n} T(y_i)\right)\left[\int \eta(\theta_2)p_2(\theta_2)\,\mathrm{d}\theta_2 - \int \eta(\theta_1)p_1(\theta_1)\,\mathrm{d}\theta_1\right] \\ - 2n\left[\int A(\theta_2)p_2(\theta_2)\,\mathrm{d}\theta_2 - \int A(\theta_1)p_1(\theta_1)\,\mathrm{d}\theta_1\right],\end{aligned} \tag{3.6}$$

where $T(y)$ is a sufficient statistic for the model parameter θ, $\eta(\theta)$ is a function of θ (called the natural parameter), $A(\theta)$ is the logarithm of a normalizing constant, and $p_1(\theta_1)$ and $p_2(\theta_2)$ are prior distributions representing different hypotheses about θ (possibly on subdimensions of θ, hence subscripts 1 and 2 on θ). The observations, $y_1, \ldots, y_n$, are independently and identically distributed, under the condition of which $\sum_{i=1}^{n} T(y_i)$ becomes a sufficient statistic made of all observations.[2]

The above form of ΔPIC, of which all the expressions of ΔPIC in Table 3.1 for binomial model testing are special cases, shows that the criterion is a linear function of the model's sufficient statistics. It also manifests that the decision bound, or the sufficient statistic normalized by the sample size n (i.e., $\frac{\sum_{i=1}^{n} T(y_i)}{n}$) solving $\Delta\text{PIC} = 0$, remains constant for all n's. Therefore, in its applications to hypothesis testing, we expect precisely the same behavior as demonstrated with binomial models to arise: There exist a range of observed statistics that cannot evidence one of the models, no matter how strong evidence is accumulated in the statistic in support of the model.

We believe that the PIC's problematic selection extends to models outside exponential families, but its specific form is uncertain. Nonetheless, some insight can be gained by considering the following relationship:

$$\begin{aligned}
\text{PIC} &= -2 \int \log[f(y|\theta)] p(\theta) \mathrm{d}\theta \\
&= -2 \int \log \left[\int f(y|\theta) p(\theta) \mathrm{d}\theta \cdot \frac{f(y|\theta)p(\theta)}{\int f(y|\theta)p(\theta)\mathrm{d}\theta} \cdot \frac{1}{p(\theta)} \right] p(\theta)\mathrm{d}\theta \\
&= -2 \int \log \left[\int f(y|\theta) p(\theta) \mathrm{d}\theta \cdot \frac{p(\theta|y)}{p(\theta)} \right] p(\theta)\mathrm{d}\theta \\
&= -2 \log \int f(y|\theta) p(\theta) \mathrm{d}\theta + 2 \int \left[\log \frac{p(\theta)}{p(\theta|y)} \right] p(\theta)\mathrm{d}\theta \\
&= -2 \log \mathrm{ML} + 2 D_{\mathrm{KL}} \left(p(\theta) || p(\theta|y) \right),
\end{aligned} \tag{3.7}$$

[2] This expression extends to the case of a vector parameter in a straightforward fashion. In such a case, the first term of ΔPIC in Equation 3.6 is replaced by a linear combination of multivariate sufficient statistics.

where $D_{\mathrm{KL}}(p(\theta)||p(\theta|y))$ is the Kullback-Leibler (KL) divergence of the posterior distribution from the prior distribution. This means that the PIC can be regarded as a two-part decomposition: (the negative logarithm of) the model's marginal likelihood and the KL divergence between prior and posterior distributions. With two models under comparison, ΔPIC becomes

$$\Delta\mathrm{PIC} = -2\log\mathrm{BF} + 2\left[D_{\mathrm{KL}}\left(p_1(\theta)||p_1(\theta|y)\right) - D_{\mathrm{KL}}\left(p_2(\theta)||p_2(\theta|y)\right)\right]. \quad (3.8)$$

This result translates that the selection outcome of ΔPIC deviates from that of $-2\log\mathrm{BF}$ depending on the difference of two KL divergence terms, which indicate the relative degree of departure of each model's posterior distribution from their priors. Typically, the posterior of a model changes from a widely dispersed prior to an increasingly peaked, limiting distribution as the sample size increases. The KL divergence in the above expressions as a function of n can be considered to measure the *rate* of such convergence. In fact, the convergence rate of Bayesian posteriors under various conditions, not just the fact of convergence, is an ongoing research topic and the KL divergence of a posterior plays a key role in such analysis (Ghosal, Ghosh, & van der Vaart, 2000; Kleijn & van der Vaart, 2006). Results relevant to the present discussion concern two distinct conditions: The model is well-specified versus mis-specified (i.e., the model contains the underlying, data-generating process vs. it does not). It has been shown that the posterior of a well-specified model converges to its limiting distribution at an optimal rate under mild regularity conditions, whereas the same rate for a mis-specified model is achieved when more stringent conditions are met. Recall that the PIC exhibits clearly counterintuitive selection behavior when the unconstrained model is well-specified yet the constrained model is mis-specified (i.e., the truth lies only in the unconstrained model). This behavior may be put in perspective when the deviation of the PIC from the BF shown above is considered together with the aforementioned findings about posterior convergence rates: the PIC can penalize the well-specified,

unconstrained model more severely than the BF does because the convergence of the unconstrained model's posterior is faster than the constrained model's, magnifying the KL divergence term in the PIC as a penalty.

General results about exact conditions in which the PIC produces selection outcomes that are distinct from the BF's for the above reason are not available. Instead, we may consider a special case in which the PIC's deviation from the BF takes a simplest form: when a point hypothesis, in which all parameters are fixed in a model, is tested against a larger model with (some of) the parameters set free. An example is the test of an equality constrained binomial model (M_c) against the unconstrained model (M_{01}) illustrated earlier (results shown in the bottom row of Figure 3.2). In this case, the KL divergence in Equation 3.8 is always zero for the constrained model (Model 2) because its prior and posterior are an identical point mass, whereas the KL divergence for the unconstrained model (Model 1) becomes positive as its posterior differs from the prior with data accumulation. This makes the difference of two KL divergences in Equation 3.8 strictly positive, leading to $\Delta\text{PIC} > -2\log\text{BF}$. Consequently, when compared to the point hypothesis, the unconstrained model tends to be favored less under the PIC than under the BF. This observation certainly extends to general cases. That is, with the PIC, an unconstrained, well-specified model can be penalized over the mis-specified, point hypothesis relative to the case with the BF.

Discussion

The present chapter concerns a newly proposed method of statistical model selection, the prior information criterion van de Schoot et al. (2012), and reports an analysis of its behavior when applied to test an inequality- or an equality-constrained hypothesis. The results show that the PIC yields puzzling outcomes of model selection, which are demonstrated in examples

of testing binomial models under various constraints, and explained using analytic derivations of the PIC in more general settings. In sum, it was found that there exists a situation in which one of the models under comparison cannot be selected by the PIC even when the data sensibly support that model, no matter how much data are collected. Specifically, it is possible that the model chosen by the PIC does not contain the underlying truth whereas the alternative model does.[3] In the same situation, the Bayes factor favors the model that receives sensible support from the data. As predicted in general cases (Schwartz, 1965), the selection behavior of the Bayes factor is consistent, favoring the correct, data-generating model as data accumulate.

Inconsistent selection behavior is in fact a large-sample property of some existing model selection criteria, most often associated with the Akaike information criterion (AIC; Akaike, 1973). One may wonder if the inconsistency of the PIC can be understood in a similar way. There are critical differences, however, in what constitutes "inconsistency" between the AIC and the PIC. The AIC is inconsistent in the sense that it is possible for a constrained model not to be recovered with a large sample when the true distribution lies both in the constrained and the alternative, encompassing model (Bozdogan, 1987). This selection is viewed as being inconsistent because, in such a case, the smaller, nested model should be considered closer to the truth than the larger model (i.e., inconsistent with respect to the order of model classes in their overall discrepancy from the true distribution). This outcome is far from being puzzling and is defended by the fact that the AIC is designed to estimate the fitted model's KL divergence from the true

[3] In fact, an example of how the PIC can lead to a problematic inference of this kind is already present (but neither observed nor discussed) in van de Schoot et al. (2012). For example, when data were simulated from a population where $\mu_1 = -1$ and $\mu_2 = 1$, the PIC preferred the equality constrained model ($\mu_1 = \mu_2$) over the unconstrained model (μ_1, μ_2), despite the fact that the unconstrained, but not the constrained model, includes the truth (see their Figure 3 and Mulder, 2014).

distribution, not the order of compared model classes in the aforementioned sense. By contrast, the PIC does not let one select a model even when the truth lies *only* in that model and not in the alternative one. Furthermore, the PIC in certain cases does not recover the true model with data of all sample sizes. To the best of our knowledge, there are no existing model selection criteria that exhibit such behavior.

Methods for testing hypotheses statistically, or selecting among competing statistical models, are important tools for scientific research. Development of a new method should be welcomed as there must not be a single, absolute criterion for such testing. The PIC would have been a valuable addition to the existing methods. However, based on our analyses reported in the current chapter, we must express reservations about its use.

Appendix

This section shows the analytic derivations of the PIC and the $-2\log \text{ML}$ applied to the binomial likelihood with four different priors: An unconstrained model, $M_{01} : 0 < \theta < 1$ with $\theta \sim \text{Beta}(\alpha, \beta)$, a one-sided and two sided inequality constrained model, $M_{0u} : 0 < \theta < u$ and $M_{lu} : l < \theta < u$, with $\theta \sim \text{TruncatedBeta}(\alpha, \beta)$, and an equality constrained model, $M_c : \theta = c$.

PIC

$$
\begin{aligned}
\text{PIC}_{\text{M}_{01}} &= -2 \int_0^1 \log\left(f(y|\theta)\right) p(\theta)\, \mathrm{d}\theta \\
&= -2 \int_0^1 \log\left(\binom{n}{y} \theta^y \left(1-\theta\right)^{n-y}\right) \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\, \mathrm{d}\theta \\
&= -2\log\binom{n}{y} - 2y \int_0^1 \log\left(\theta\right) \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\, \mathrm{d}\theta \\
&\quad - 2(n-y) \int_0^1 \log\left(1-\theta\right) \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\, \mathrm{d}\theta \\
&= -2\log\binom{n}{y} - 2y\left(\psi(\alpha) - \psi(\alpha+\beta)\right) - 2(n-y)\left(\psi(\beta) - \psi(\alpha+\beta)\right) \\
&= -2\log\binom{n}{y} - 2y\psi(\alpha) - 2n\psi(\beta) + 2n\psi(\alpha+\beta) + 2y\psi(\beta) \\
&= -2\log\binom{n}{y} - 2y\left(\psi(\alpha) - \psi(\beta)\right) - 2n\left(\psi(\beta) - \psi(\alpha+\beta)\right),
\end{aligned}
\tag{3.9}
$$

where $B\left(\cdot,\cdot\right)$ is the beta function and $\psi\left(\cdot\right)$ is the digamma function.

Setting $\alpha = \beta = 1$ yields

$$
\text{PIC}_{\text{M}_{01}} = -2\log\binom{n}{y} + 2n, \tag{3.10}
$$

since $\psi(1) - \psi(2) = -1$.

$$\begin{aligned}
\mathrm{PIC}_{\mathrm{M0u}} &= -2\int_0^u \log\left(\binom{n}{y}\theta^y(1-\theta)^{n-y}\right)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B_u(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log\binom{n}{y} - 2n\int_0^u \log(1-\theta)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B_u(\alpha,\beta)}\,\mathrm{d}\theta \\
&\quad - 2y\int_0^u \log(\theta)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B_u(\alpha,\beta)}\,\mathrm{d}\theta + 2y\int_0^u \log(1-\theta)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B_u(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log\binom{n}{y} - 2y\frac{\partial_\alpha B_u(\alpha,\beta) - \partial_\beta B_u(\alpha,\beta)}{B_u(\alpha,\beta)} - 2n\frac{\partial_\beta B_u(\alpha,\beta)}{B_u(\alpha,\beta)},
\end{aligned} \tag{3.11}$$

where $B_u(\cdot,\cdot)$ is the incomplete beta function, ∂_α denotes the partial derivative with respect to α and ∂_β denotes the partial derivative with respect to β. The partial derivatives of the incomplete beta function, $\partial_\alpha B_u(\alpha,\beta)$ and $\partial_\beta B_u(\alpha,\beta)$ have a closed form expression in terms of generalized hypergeometric functions, $F(a_1,\ldots,a_p;b_1,\ldots,b_q;z)$:

$$\partial_\alpha B_u(\alpha,\beta) = \log(u)B_u(\alpha,\beta) - \frac{u^\alpha}{\alpha^2}F(\alpha,\alpha,1-\beta;\alpha+1,\alpha+1;u). \tag{3.12}$$

$$\begin{aligned}
\partial_\beta B_u(\alpha,\beta) &= \frac{(1-u)^\beta}{\beta^2}F(\beta,\beta,1-\alpha;\beta+1,\beta+1;1-u) \\
&\quad - \log(1-u)B_{1-u}(\beta,\alpha) + \left(\psi(\beta) - \psi(\alpha+\beta)\right)B(\alpha,\beta).
\end{aligned} \tag{3.13}$$

Setting $\alpha = \beta = 1$ yields

$$\mathrm{PIC}_{\mathrm{M0u}} = -2\log\binom{n}{y} - 2y\left(\log(u) - \frac{u-1}{u}\log(1-u)\right) - 2n\left(\frac{u-1}{u}\log(1-u) - 1\right), \tag{3.14}$$

since $B_u(1,1) = u$; $\partial_\alpha B_u(1,1) = u\,(\log(u)-1)$ and $\partial_\beta B_u(1,1) = (u-1)\log(1-u) - u$. As desired, if $u = 1$, Equation (3.11) reduces to Equation (3.9) and Equation (3.14) reduces to Equation (3.10).

$$\begin{aligned}
\mathrm{PIC}_{\mathrm{M_{lu}}} &= -2\int_l^u \log\left(\binom{n}{y}\theta^y(1-\theta)^{n-y}\right)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{D_{ul}(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log\binom{n}{y} - 2n\int_l^u \log(1-\theta)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{D_{ul}(\alpha,\beta)}\,\mathrm{d}\theta \\
&- 2y\int_l^u \log(\theta)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{D_{ul}(\alpha,\beta)}\,\mathrm{d}\theta + 2y\int_l^u \log(1-\theta)\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{D_{ul}(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log\binom{n}{y} - 2y\frac{\partial_\alpha D_{ul}(\alpha,\beta) - \partial_\beta D_{ul}(\alpha,\beta)}{D_{ul}(\alpha,\beta)} - 2n\frac{\partial_\beta D_{ul}(\alpha,\beta)}{D_{ul}(\alpha,\beta)},
\end{aligned} \tag{3.15}$$

where we define the difference between two incomplete beta functions as follows: $D_{ul}(\alpha,\beta) = B_u(\alpha,\beta) - B_l(\alpha,\beta) = \int_l^u \theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta$.

Setting $\alpha = \beta = 1$ yields

$$\begin{aligned}
\mathrm{PIC}_{\mathrm{M_{lu}}} &= -2\log\binom{n}{y} - 2y\frac{-(1-l)\log(1-l) + u\log(u) + (1-u)\log(1-u) - l\log(l)}{u-l} \\
&- 2n\frac{(1-l)(\log(1-l)-1) - (1-u)(\log(1-u)-1)}{u-l},
\end{aligned} \tag{3.16}$$

since $\partial_\alpha D_{ul}(1,1) = -l(\log(l)-1) + u(\log(u)-1)$ and $\partial_\beta D_{ul}(1,1) = (1-l)\,(\log(1-l)-1) - (1-u)\,(\log(1-u)-1)1$. As desired, if $l = 0$, Equation (3.15) reduces to Equation (3.11) and Equation (3.16) reduces to Equation (3.14).

$$\begin{aligned}
\mathrm{PIC}_{\mathrm{M_c}} &= -2\log\binom{n}{y}c^y\,(1-c)^{n-y} \\
&= -2\log\binom{n}{y} - 2y\log(c) - 2(n-y)\log(1-c) \\
&= -2\log\binom{n}{y} + 2y\big(-\log(c) + \log(1-c)\big) - 2n\log(1-c) \\
&= -2\log\binom{n}{y} - 2y\log\left(\frac{c}{1-c}\right) - 2n\log\left(1-c\right).
\end{aligned} \tag{3.17}$$

-2 log ML

$$\begin{aligned}
-2\log \mathrm{ML}_{\mathrm{M}_{01}} &= -2\log \int_0^1 f(y|\theta)p(\theta)\,\mathrm{d}\theta \\
&= -2\log \int_0^1 \binom{n}{y}\theta^y\,(1-\theta)^{n-y}\,\frac{\theta^{a-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log \frac{\binom{n}{y}}{B(\alpha,\beta)}\int_0^1 \theta^{y+\alpha-1}\,(1-\theta)^{n-y+\beta-1}\,\mathrm{d}\theta \\
&= -2\log \frac{\binom{n}{y}B(y+\alpha,n-y+\beta)}{B(\alpha,\beta)}.
\end{aligned} \tag{3.18}$$

Setting $\alpha = \beta = 1$, yields

$$\begin{aligned}
-2\log \mathrm{ML}_{\mathrm{M}_{01}} &= -2\log \binom{n}{y}B(y+1,n-y+1) \\
&= -2\log \frac{1}{n+1},
\end{aligned} \tag{3.19}$$

since $B(m+1,n+1) = \frac{m!n!}{(m+n+1)!}$.

$$\begin{aligned}
-2\log \mathrm{ML}_{\mathrm{M}_{0\mathrm{u}}} &= -2\log \int_0^u \binom{n}{y}\theta^y\,(1-\theta)^{n-y}\,\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B_u(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log \frac{\binom{n}{y}}{B_u(\alpha,\beta)}\int_0^u \theta^{y+\alpha-1}\,(1-\theta)^{n-y+\beta-1}\,\mathrm{d}\theta \\
&= -2\log \frac{\binom{n}{y}B_u(y+\alpha,n-y+\beta)}{B_u(\alpha,\beta)}.
\end{aligned} \tag{3.20}$$

Setting $\alpha = \beta = 1$, yields

$$-2\log \mathrm{ML}_{\mathrm{M}_{0\mathrm{u}}} = -2\log \binom{n}{y}\frac{B_u\,(y+1,n-y+1)}{u}, \tag{3.21}$$

since $B_u(1,1) = u$. As desired, if $u = 1$, Equation (3.20) reduces to equation (3.18) and Equation (3.21) reduces to Equation (3.19).

$$\begin{aligned}
-2\log \mathrm{ML}_{\mathrm{M}_{\mathrm{lu}}} &= -2\log \int_l^u \binom{n}{y}\theta^y\,(1-\theta)^{n-y}\,\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{D_{ul}(\alpha,\beta)}\,\mathrm{d}\theta \\
&= -2\log \frac{\binom{n}{y}}{D_{ul}(\alpha,\beta)}\int_l^u \theta^{y+\alpha-1}\,(1-\theta)^{n-y+\beta-1}\,\mathrm{d}\theta \\
&= -2\log \frac{\binom{n}{y}D_{ul}(y+\alpha,n-y+\beta)}{D_{ul}(\alpha,\beta)}.
\end{aligned} \tag{3.22}$$

Setting $\alpha = \beta = 1$, yields

$$-2\log \mathrm{ML}_{\mathrm{M_{ul}}} = -2\log \binom{n}{y} \frac{D_{ul}\left(y+1, n-y+1\right)}{u-l}. \tag{3.23}$$

As desired, if $l = 0$, Equation (3.22) reduces to Equation (3.20) and Equation (3.23) reduces to Equation (3.21).

$$-2\log \mathrm{ML}_{\mathrm{M_c}} = -2\log \binom{n}{y} - 2y\log\left(\frac{c}{1-c}\right) - 2n\log(1-c). \tag{3.24}$$

Chapter 4

Using parameter space partitioning to evaluate a model's qualitative fit

Introduction

Parameter space partitioning (PSP; Pitt et al., 2006) is a versatile tool for model analysis, focusing on qualitative model behavior. It makes abstraction of the quantitative details in model predictions and focuses on the more general qualitative data patterns, such as ordinal relations. PSP is implemented using a Markov chain Monte Carlo (Gilks, Richardson, & Spiegelhalter, 1996) search algorithm that samples points in the parameter space and evaluates the qualitative model behavior in every selected set of parameter values. The entire parameter space gets partitioned into regions that correspond to the different qualitative data patterns the model can generate. All points belonging to the same region generate an identical qualitative data pattern. Additionally, every region is provided with an estimate of the volume it occupies in the space, reflecting the representativeness of the corresponding data pattern to the model's behavior.

Pitt et al. (2006) showed that PSP can be used for a number of purposes. First, they applied PSP to evaluate the qualitative behavior of the human category learning model ALCOVE (Kruschke, 1992) in the Shepard,

Hovland, and Jenkins (1961) task, where six different category structures are learned. The qualitative pattern of interest was the ordering of the ease with which the different structures were learned. Pitt et al. (2006) evaluated ALCOVE's behavior by inspecting different aspects of its partitioned parameter space. A count of the number of generated data patterns revealed that the variety of behavioral patterns predicted by ALCOVE was quite small, and included the empirical pattern. Pitt et al. (2006) also more closely inspected the most representative data patterns (i.e., patterns which occupied the most volume in the parameter space) and found that the orderings of the different structures in these patterns were overall quite similar to the ordering in the empirical pattern. These observations suggested that ALCOVE can account for the qualitative structure of the empirical data.

In a second application, Pitt et al. (2006) applied PSP to compare the qualitative model behavior of two localist connectionist models of speech perception, TRACE (McClelland & Elman, 1986) and Merge (Norris, McQueen, & Cutler, 2000), in two different experimental settings. The models' partitioned parameter spaces were again compared on different aspects. Pitt et al. (2006) compared the flexibility of both models by counting the number of data patterns that both models generated, and evaluated the model mimicry by studying the number of overlapping and unique data patterns (see also Hawkins, Brown, Steyvers, & Wagenmakers, 2012; Steingroever, Wetzels, & Wagenmakers, 2013 for other applications of PSP to examine model flexibility and mimicry). Further, they compared how representative the empirical pattern was to the behavior of each model by looking at the volumes of the regions in the parameter spaces that were occupied by the empirical pattern. The data patterns were also compared by means of histograms presenting the number of mismatches between the data patterns and the empirical one, and the most representative data patterns were inspected with more scrutiny, relating the volumes of their regions to the

number of mismatches. Overall, these analyses suggested that TRACE and Merge show a lot of similarities in their qualitative behavior.

These and other applications make clear that PSP is an extremely rich source that provides a great amount of useful information about a model's qualitative behavior: Does the model generate the empirical data pattern? How representative is the empirical data pattern to the model? Does the model generate many other data patterns besides the empirical one? Do the other generated data patterns deviate a lot from the empirical one? How representative are these other data patterns to the model? This richness of information, however, has also a drawback. Particularly when PSP is used for different models and one wishes to select one of the competing models as the one that performs best in capturing the empirical qualitative pattern, it is not at all straightforward how the different pieces of information provided by a PSP analysis should be meaningfully combined. Especially when the models can generate many data patterns (e.g., thousands) and the data patterns are large (e.g., rankings of more than 10 stimuli), it can become a cumbersome task to find out which model provides the best qualitative account of the data.

In this chapter, we propose a convenient way of summarizing the outcome of a PSP analysis into a single number useful for model selection, referred to as PSP fit. This reflects a model's ability to account for the empirical qualitative pattern. An additional goal of this chapter is to evaluate the performance of PSP based model selection using an extensive simulation study.

The remainder of the chapter is organized as follows. First, we explain how PSP can be used to select among models in five steps. Next, we show four application examples where we use PSP fit to select among two competing models in the field of category learning. Finally, we evaluate the ability of PSP fit as a model selection method using a model recovery study.

PSP based model selection in five steps

To explain PSP based model selection, we use an example of two hypothetical models, each with two parameters, denoted by θ_1 and θ_2, applied to a response time task with five conditions. We illustrate how PSP can be used as a tool for model selection in five steps.

Step 1: Define qualitative data pattern

The first step is to define the qualitative data pattern, depending on the interest of the researcher. In the current example, we assume the main interest of the researcher is capturing the ordering of the mean response times across five conditions, and thus a data pattern is defined as the rank order of mean response times across all conditions.

Step 2: Run PSP

The second step involves running PSP using the data pattern definition from step 1 (for this we use Matlab code available on `http://faculty.psy.ohio-state.edu/myung/personal/psp.html`). Figure 4.1 illustrates the partitioned parameter spaces of the two models, showing all different data patterns (in this case, rank orders) they can generate across their parameter space. Model 1's parameter space is partitioned into seven regions, corresponding to seven different rank orders of the five conditions. In each predicted pattern, the first number reflects the rank of condition 1, the second number reflects the rank of condition 2, and so forth. Rank 1 reflects the fastest and rank 5 reflects the slowest response time. For example, data pattern (5, 1, 2, 3, 4) indicates that condition 1 evoked the slowest response time, condition 2 evoked the fastest response time, and condition 3 to 5 evoked the second, third and fourth fastest response times, respectively. This data pattern occupies about one third of the space and is the

most representative of the model behavior. Being assigned a much smaller volume of 5%, data pattern (3, 5, 2, 4, 1) is less relevant. Model 2 generates five different rank orders, the most important ones being (3, 4, 2, 5, 1), (3, 5, 1, 4, 2) and (2, 5, 3, 4, 1).

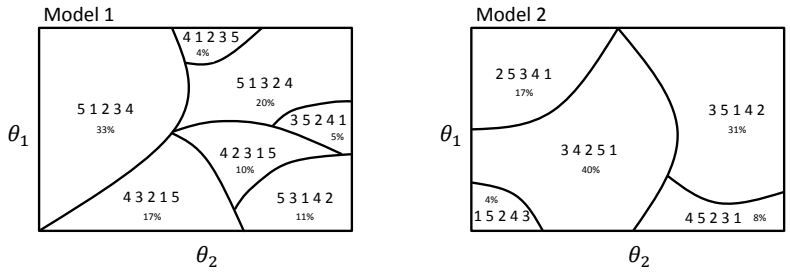


Figure 4.1: *The partitioned parameter spaces of two hypothetical two-parameter models in an experiment with five conditions. Model 1 generates seven distinct data patterns, whereas model 2 generates five different qualitative patterns. Every region depicts its corresponding data pattern, and the percentage of the volume that it occupies in the model's parameter space. This Figure is inspired by Figure 2 in Pitt et al. (2006).*

Step 3: Assess model distinguishability

The third step involves assessing the distinguishability of the models under consideration. Assessing distinguishability is crucial, since no model selection method can be reasonably expected to disentangle two models if they make predictions that are very much alike (Navarro, Pitt, & Myung, 2004; Pitt, Kim, & Myung, 2003). An important benefit of using PSP to select models is that distinguishability can be assessed based on the PSP output, precluding having to perform additional simulations. A comparison of the data patterns generated by model 1 and model 2 in Figure 4.1 shows that they have no data patterns in common, so the models can be dinstinguished.

Step 4: Collect data

Having ascertained distinguishability, it makes sense to collect empirical data. In this example, we assume that the observed response times (in ms) across the five conditions are (932, 1196, 745, 1071, 540). The qualitative representation of these data, using the data pattern definition from step 1, is (3, 5, 2, 4, 1).

Step 5: Compute PSP fit

The fifth and final step involves comparing the competing models on their ability to describe the observed data pattern. A first intuitive way of bringing the observed data pattern into contact with the predicted data patterns involves looking for the observed pattern among the predicted data patterns. The partitioned parameter spaces in Figure 4.1 show that model 1 is able to generate the empirical data pattern, whereas model 2 is not. From this perspective, one could conclude that model 1 describes the observed qualitative trend better and should be preferred over model 2. The major problem with this intuitive approach is that it is highly prone to over-fitting. Over-fitting occurs when a model does not only fit variation due to the underlying process, but also captures random variation. Over-fitting is usually caused by a model being too complex: An overly complex model provides a good fit to the data due to its flexibility to capture random error in the data and does not accurately reflect the underlying process. To reduce the risk of over-fitting, one should not only assess a model's ability to fit the data, but also take into account the price that has to be paid to achieve this fit, in terms of the model's complexity (Myung, 2000; Pitt & Myung, 2002; Roberts & Pashler, 2000). Choosing the model that can describe the data well without being too complex is an implementation of Occam's razor.

One common solution to avoid over-fitting involves considering the average fit of a model, across the entire domain of parameter values (Myung

& Pitt, 1997; Pitt, Myung, & Zhang, 2002). Model complexity is taken into account in this way, in the sense that a model is penalized for being able to fit data patterns that heavily diverge from the observed data. Applied to PSP based model selection, evaluating average model fit implies an approach that evaluates the global correspondence between the empirical data pattern and *all* the data patterns generated by the model. As explained in the introduction section, this comes down to weighing different aspects of a model's partitioned parameter space, such as how much the different data patterns deviate from the empirical one and how representative they are to the model's behavior.

The partitioned parameter spaces in Figure 4.1 show that although model 1 can generate the empirical pattern, the region of model 1 corresponding to this pattern is not very representative for the model, as it only occupies a relatively small part of the parameter space. Moreover, the other data patterns generated by model 1 are rather different from the empirical one. For example, whereas the fifth and second condition have the highest and lowest empirical ranking respectively, this seems to be almost opposite in most data patterns generated by model 1. In contrast, the majority of the data patterns produced by model 2 show great similarities with the empirical one. Almost all pairwise relations found in the empirical data are preserved in model 2's patterns, except for some minor switches, with the most representative patterns showing the smallest number of switches. Thus, after inspecting all generated data patterns and comparing them with the empirical pattern, it seems reasonable to conclude that globally, model 2 captures the observed qualitative trend better than model 1 does.

Whereas evaluating and comparing both models' partitioned parameter spaces in Figure 4.1 is easy, in our experience, models often generate a large number of data patterns, each occupying tiny fractions of the parameter space, making it difficult to get a clear view on the models' global behavior.

Especially when competing models make similar predictions, identifying the model that provides the best overall description of the empirical pattern can be challenging. Therefore, we propose to formally summarize a model's average PSP fit as follows, allowing an easy way of selecting between models based on PSP:

$$\text{PSP fit} = \frac{\sum_{i=1}^{n} d\left(DP^{\text{Emp}}, DP_i^{\text{Model}}\right) \cdot V_i}{d\left(DP^{\text{Emp}}, DP^{\text{Max}}\right)}, \tag{4.1}$$

where n indicates the number of data patterns generated by the model and d indicates the distance between two data patterns. DP^{Emp}, DP_i^{Model} and DP^{Max} correspond to the empirical data pattern, the ith data pattern generated by the model, and the theoretically most distant data pattern from the empirical one, respectively. Further, V_i is the proportion of the volume that data pattern i occupies in the model's parameter space.

The numerator is the summed distance (which is a positive value) between the empirical pattern and each data pattern generated by the model, including implausible ones. Each term is weighted by the proportion of the volume of the corresponding data pattern's region. It reflects the weighted average distance between a model's data patterns and the empirical data pattern. The weighted distance is then divided by its maximum value, which is equal to the distance between the empirical pattern and the theoretically most distant data pattern from the empirical one.[1] In this way, the average weighted fit is scaled to fall in the range $[0, 1]$, with a lower value indicating a lower distance between model and data, and thus a better global fit. The minimal value of 0 is obtained if the model generates a single data pattern only, corresponding to the empirical pattern. As soon as other patterns are

[1] The theoretically most distant data pattern from the empirical one does not necessarily need to be one of the data patterns generated by the model, so the distance between the empirical data pattern and the theoretically most distant data pattern will always be larger than zero.

generated, the value of PSP fit increases. The maximal value of 1 is obtained if the model generates a single data pattern only, corresponding to the pattern most distant from the empirical one.

The exact implementation of the distance between two data patterns depends on the way patterns are defined in step 1. In the current example, as well as in the applications below, we focus on rank order patterns, so we need a measure that is suitable to assess the distance between rankings. To select from the wealth of ranking distances that are available (see e.g., Marden, 1995, for an overview), we restricted ourselves to measures that can be applied to partial rankings. Such rankings have ties, meaning there are certain items that have the same ranking, and can reasonably be expected to occur in psychological experiments (e.g., see the Shepard et al. (1961) task described earlier). Fagin, Kumar, Mahdian, Sivakumar, and Vee (2006) introduced four distance metrics that can be used to compare such partial rankings, obtained by generalizations of the well-known Kendall tau distance[2] (Kendall, 1938, 1975) and Spearman's footrule distance (Spearman, 1904; Diaconis & Graham, 1977). We will use the most easily implementable metric, which is Spearman's footrule distance generalized to partial rankings by assigning "midrank" values to the tied items or stimuli. This means that ties are replaced by the average of all tied positions in the ranking. For example, if the first and third stimulus in our example given above (see step 4) are tied, the empirical data pattern would be (2.5, 5, 2.5, 4, 1) instead of (3, 5, 2, 4, 1). Spearman's footrule distance between two (partial) rankings can then be computed as follows: $d\left(DP_1, DP_2\right) = \sum_{j=1}^{m} \left|DP_{1j} - DP_{2j}\right|$, with DP_{1j} and DP_{2j} the rank position of the jth element in the first and second data pattern, respectively.

Using this distance, the PSP fit of model 1 and model 2 equals 0.81 and 0.17, respectively (see Appendix A for computational details), indicating

[2]Not to be confused with Kendall's tau coefficient.

that model 2 provides a better overall description of the qualitative trend in the data than model 1.

By construction, PSP fit is not only sensitive to which patterns are generated, but also to the proportion of the volumes of the regions these patterns occupy in the parameter space. If, for example, model 1 would generate the data patterns shown in Figure 4.1, but with the volume of the empirical pattern's region equal to 88% and the volumes of the other regions equal to 2%, the PSP fit of model 1 would drop to 0.10 so that model 1 is preferred.

Application Example

In this section, we demonstrate PSP based model selection in the domain of category learning. We consider two category learning models that have a long history of being contrasted (see, e.g., Vanpaemel & Storms, 2010, for an overview): the Generalized Context Model (GCM: Nosofsky, 1986) and the Multiplicative Prototype Model (MPM: Minda & Smith, 2011; Nosofsky, 1987; Reed, 1972). These models reflect two opposing theoretical accounts of categorization, according to which humans classify items in a certain category by comparing them to category exemplars or to a category prototype, respectively. Formal details of the models used in this application are provided in Appendix B[3].

In a typical category learning task, participants are asked to classify

[3]Our comparison between the GCM and the MPM serves as a demonstration of PSP fit as a model selection method and not as a way to draw substantial conclusions concerning the exemplar versus prototype accounts of category learning, as more recent model versions could be considered for this latter purpose (Ashby & Maddox, 1993; Vanpaemel, 2016). Further, the four data sets considered here are not intended to be representative for the vast collection of category learning data sets available, which can differ in many meaningful characteristics, including the category structures to be learned, the type of stimuli, the exact instructions, the amount of training, the type of feedback, and so on.

previously unseen stimuli in category A or B. During the training phase, participants are trained to classify a subset of the stimuli (i.e., the assigned stimuli) in one of both categories, based on corrective feedback. During the test phase, both the assigned and the unassigned stimuli have to be categorized. In this application example, we use four category structures from Nosofsky (1986), referred to as criss-cross, interior-exterior, diagonal, and dimensional. They are based on the same set of 16 stimuli (semicircles), varying on two dimensions (size of the semicircle and angle of orientation of a radial line drawn from the center of the semicircle to the rim). The four structures are depicted schematically in the upper left panel of Figures 4.2 to 4.5. Every cell in these grids schematically represents one of the 16 stimuli. Rows and columns reflect the four levels of the two dimensions on which the stimuli vary.

As a start, let us consider the criss-cross structure and apply the five steps of the PSP based model selection procedure that we have outlined above. The first step involves deciding on the data pattern of interest. In this application, we focus on the ordinal relations between the category A responses for all stimuli (e.g., stimulus 1 was classified more often in category A than stimulus 2, but less often than stimulus 3), so we define a qualitative data pattern as the rank order of the percentage of category A responses (rounded to the nearest integer) for all stimuli. The stimulus that was assigned most often to category A gets a rank of 1, while the stimulus that was assigned most often to category B gets a rank of 16. Stimuli with equal category A response percentages get a "midrank" value.

Second, PSP is run. PSP revealed 8426 different data patterns generated by the GCM and 2142 data patterns generated by the MPM. The interested reader can find details of the PSP procedure in Appendix C.

The third step involves assessing model distinguishability. The two bottom rows in Figure 4.2 show a graphical representation of the five data

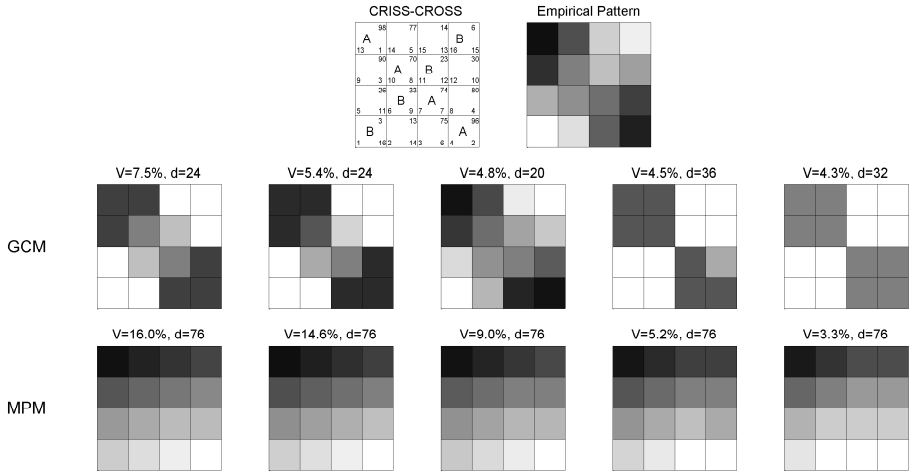


Figure 4.2: *The top left panel shows the criss-cross category structure and the data, as collected by Nosofsky (1986). Rows and columns reflect the four levels of the two dimensions on which the 16 stimuli vary. Cells that are marked with an A represent stimuli assigned to category A, whereas cells with a B represent stimuli assigned to category B. Cells without a letter are unassigned stimuli. Stimulus numbers are depicted in the lower left corner of each cell. The observed category A response percentage for each stimulus is depicted in the upper right corner of each cell. The corresponding rank position in the empirical data pattern is depicted in the lower right corner of each cell, with 1 as the highest and 16 as the lowest possible rank position. The top right panel shows a visualization of the empirical data pattern, where each cell in the grid represents one of the 16 stimuli to be categorized. The darker the cell, the higher the rank order position of that stimulus in the concerning pattern. The degree of darkness hence reflects the preference for category A. The middle row depicts a visualization of the five most representative predicted patterns of the GCM, and the bottom row depicts the five most representative data patterns of the MPM, using the same color coding conventions as in the visualization of the empirical data pattern. Above each predicted data pattern, the percentage of the volume (V) of the pattern and the distance (d) between the data pattern and the empirical pattern is shown.*

patterns with the largest volumes in each model's partitioned parameter space. The gradient of each cell reflects the rank position of the corre-

sponding stimulus, with a darker gradient reflecting a higher ranking (i.e., more category A responses). This visualization provides an easy insight into the most representative qualitative model behavior and allows a quick way to roughly check model distinguishability for the largest patterns. Clearly, both models generate different types of data patterns. In the GCM patterns, stimuli that simultaneously have a low value on one and a high value on the other dimension show high rankings (dark gradient, reflecting more category A responses), whereas low or high values on both dimensions result in low rankings (light gradient, reflecting more category B responses). In contrast, rankings in the MPM patterns consistently increase with higher values on the vertical dimension and lower values on the horizontal dimension[4]. Thus, the models seem to be distinguishable.

Convinced of the models' distinguishability, it makes sense to collect data in a fourth step; or in this case, consider the data already collected by Nosofsky (1986). In this application, we will use the data obtained from Participant 1. Using the data pattern definition from step 1, the empirical qualitative data (i.e., the rankings of all stimuli) are shown in the lower right corners of the upper left panel of Figure 4.2; the right panel shows a graphical representation of this qualitative data pattern. For completeness, the quantitative data — the observed percentages of category A responses, rounded to the nearest integer — are shown in the upper right corners of the cells, but these are not used in the application.

Now, the PSP fit between the empirical data pattern and the model-based data patterns can be assessed, using Equation (4.1). Since we defined

[4]From the schematic representation of this category structure, these asymmetric data patterns generated by the MPM are surprising, as it seems that the prototypes of A and B have the same location (namely, in the middle of the grid). However, the model uses the psychological space of Participant 1 derived from identification data (see Figure 5 in Nosofsky, 1986), in which the stimuli are not represented as symmetrically as in the grid representation.

a data pattern as a ranking, we use Spearman's footrule distance to calculate the distances in the PSP fit, like in our introductory example. A visual comparison of the largest predicted data patterns and the empirical data pattern in Figure 4.2 suggests that the predictions of the GCM provide the best qualitative fit. Calculation of the PSP fit quantifies this observation and extends it to all 8,426 and 2,142 patterns produced by each model. The PSP fits of the GCM and MPM are 0.17 and 0.59, respectively (see Table 4.1). Since a lower value reflects a better global correspondence between the empirical data pattern and the model, these values imply that the GCM captures the observed qualitative pattern better than the MPM.

Table 4.1: *Values of the PSP fit of the GCM and the MPM in the four category structures from Nosofsky (1986). The PSP fits in the dimensional structure are only displayed for completeness, as this structure allows very poor model distinguishability.*

	Model	
Structure	GCM	MPM
Criss-cross	0.17	0.59
Interior-exterior	0.36	0.65
Diagonal	0.13	0.30
Dimensional	(0.19)	(0.26)

Turning to the interior-exterior structure, Figure 4.3 shows the five largest data patterns generated by the GCM and MPM after running PSP (see Appendix C for more details). In total, 14,848 data patterns were generated by the GCM and 4,444 data patterns by the MPM. Visual inspection of the bottom rows of Figure 4.3 reveals that both models again produce very distinguishable qualitative patterns, so it makes sense to collect data to select between both models. The five primary patterns of the GCM reflect the empirical pattern more than MPM's primary patterns do. Calculation

of the PSP fits confirms that this finding generalizes to the global qualitative behavior of both models: the PSP fit equals 0.36 for the GCM and 0.65 for the MPM. Overall, GCM's performance in predicting the empirical trend in the interior-exterior structure is better than that of the MPM.

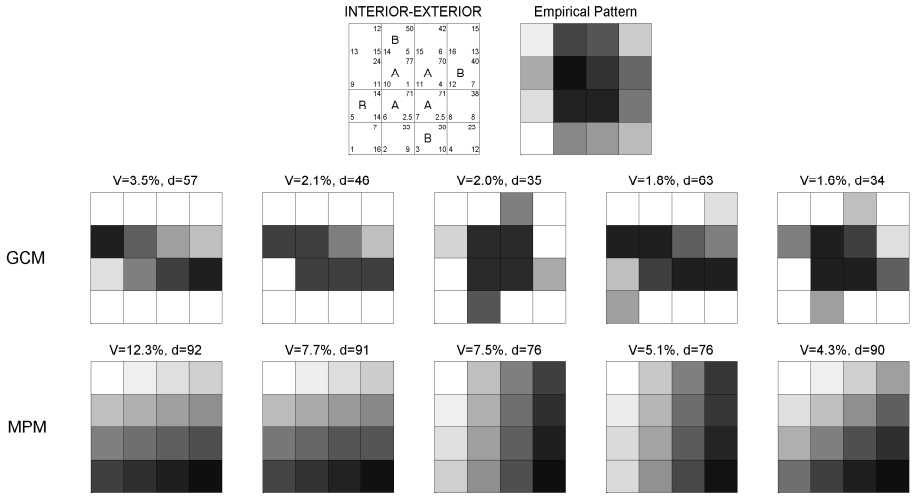


Figure 4.3: *Visualisation of the empirical and predicted data patterns in the interior-exterior structure. See Figure 4.2 for further information.*

Running PSP on the diagonal structure resulted in 3,902 data patterns generated by the GCM and 6,222 data patterns generated by the MPM (see Appendix C for more details). The five most representative patterns of each model are shown in Figure 4.4. Although these patterns are less dissimilar than in the previous structures, they are still distinguishable from each other. Whereas the stimuli's rankings in the GCM patterns are determined by their position on both dimensions together, the MPM mainly focuses on the stimuli's position on the horizontal dimension. Given the distuinguishability, there is hope that both models can be told apart, so it is worth the effort to collect empirical data. The primary GCM patterns seem to resemble the empirical pattern more than the primary MPM patterns do. Indeed, calculation of the PSP fit leads to the selection of the GCM (PSP fit = 0.13) over the MPM (PSP fit = 0.30).

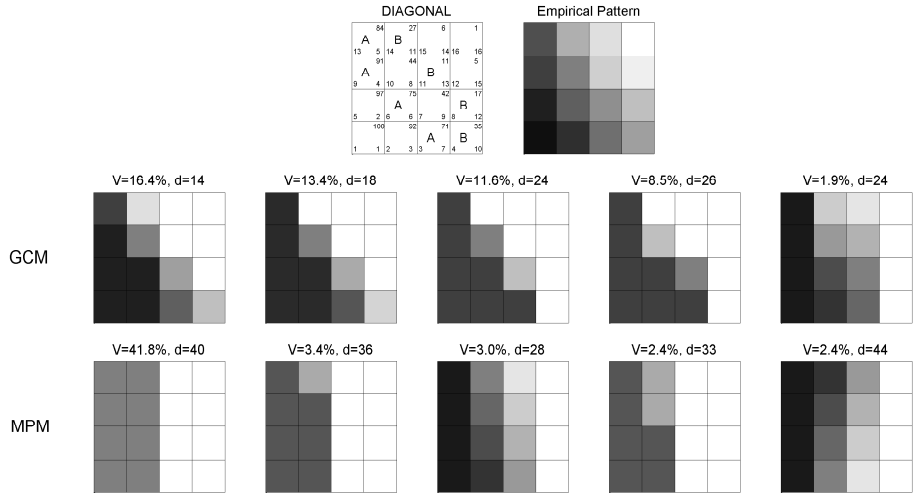


Figure 4.4: *Visualisation of the empirical and predicted data patterns in the diagonal structure. See Figure 4.2 for further information.*

Finally, we consider the dimensional structure. Running PSP resulted in 6,702 data patterns generated by the GCM and 5,023 data patterns generated by the MPM (see Appendix C for more details). Figure 4.5 shows the models' five most representative data patterns. In this structure, the GCM and the MPM seem to closely mimic each other's behavior: the two most representative data patterns of both models, covering a rather large area in the parameter spaces, are identical. This suggests that the experimental design of the dimensional structure does not elicit sufficiently informative data to distinguish between the GCM and MPM and thus is not appropriate for disentangling these models. In this light, it seems futile to try to select between these versions of the models in this design, and one should not even bother collecting data. However, given that the data are already collected by Nosofsky (1986), we show them in Figure 4.5 and we included the PSP fits in Table 4.1 for completeness.

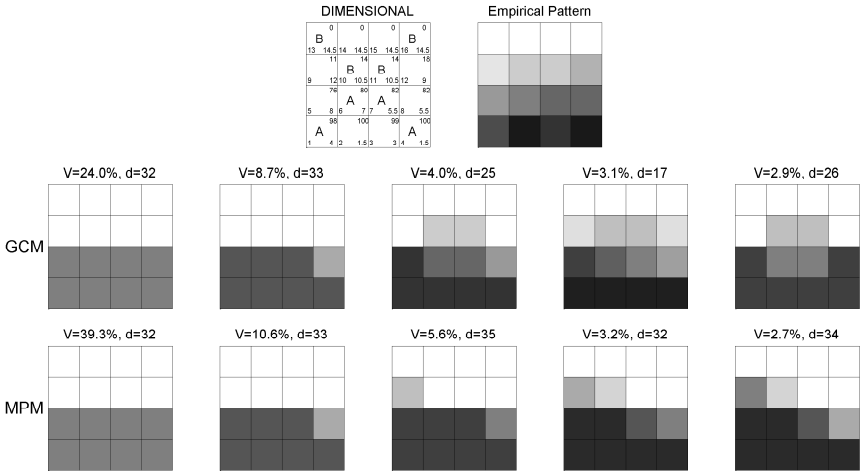


Figure 4.5: *Visualisation of the empirical and predicted data patterns in the dimensional structure. See Figure 4.2 for further information.*

Sensitivity analysis

Like all aspects of data analysis and modeling, PSP, and thus PSP fit, comes with several more or less arbitrary choices or researcher degrees of freedom (Simmons et al., 2011). First of all, there are various potential definitions of data patterns. For example, instead of considering the ranking of all stimuli, one could be interested in the pattern of A and B responses to the different stimuli (e.g., AABBA). Further, there is some freedom in choosing which stimuli to consider. Whereas we looked at all 16 stimuli that were tested, other researchers may be interested in responses to, for example, the 8 unassigned stimuli only. Another researcher degree of freedom inherent to PSP concerns the constraints on the parameter space. Although some model parameters can have natural boundaries (e.g., the attention weight parameter w is restricted to $[0, 1]$), other parameters are unbounded (e.g., the sensitivity parameter c), so researchers have to constrain the parameter range based on what they consider appropriate. One brute but informative way of dealing with these researchers degrees of freedom is performing a

sensitivity analysis, to evaluate whether and to what extent conclusions are robust across different choices.

In our sensitivity analysis, we evaluated the robustness of the above conclusions across different choices for (1) the data pattern definition, (2) the considered stimuli, and (3) the parameter range. We considered two different choices for data pattern definition: besides the ranking definition used earlier, we also considered a nominal data pattern definition, where stimuli with more or less category A than category B responses were assigned response A or response B, respectively (e.g., AAABBAAB). To assess the distances between these data patterns for the calculation of PSP fit, we used the number of mismatches, or the Hamming distance. For example, the distance between data pattern AAAABBBB and BBBBBBBB equals 4. Concerning the considered stimuli, we examined two different choices: all 16 stimuli (as earlier) and the 8 unassigned stimuli. Finally, we considered three different choices for the parameter range of c: $[0, 20]$ (as earlier), $[0, 10]$ and $[0, 5]$. In our sensitivity analysis, we combined all different choices, resulting in 12 ($2 \times 2 \times 3$) different conditions, and we calculated the PSP fits of the GCM and the MPM in each condition, for all four category learning structures, resulting in 12 tables like Table 4.1. As shown in Appendix D, the results are very similar. The main difference across the different conditions was that the GCM and the MPM became indistinguishable in the diagonal structure when only the unassigned stimuli were considered. Aside from that, the model selection conclusions were robust across the different choices for data patterns, considered stimuli and parameter ranges: whenever the models were distinguishable, the GCM provided better PSP fits than the MPM, in all category structures, across all 12 conditions.

Evaluation of PSP fit as a model selection tool

To evaluate the performance of PSP fit as a basis for selecting between models, we conducted an extensive model recovery study (see e.g., Navarro et al., 2004; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). In this study, we evaluate the ability of PSP fit to recover the data-generating model. More specifically, we simulate data from both the GCM and the MPM, and assess the percentage of time in which the model selected by PSP fit corresponds to the model that generated the data.

Using the four category structures from Nosofsky (1986) as used in the application, $80,000$ artificial data sets were generated. For each of the four structures, $10,000$ sets of parameter values of c (with range $[0, 20]$) and w (with range $[0, 1]$) were sampled from a uniform distribution on the entire parameter space. For each set of parameter values, the probability of classifying each of the 16 stimuli in category A was computed, both according to the GCM and to the MPM. Based on these probabilities, responses were generated using the binomial distribution with the sample size that had been adopted in the concerning experiment. These responses were subsequently converted into rank order data patterns by assigning a rank position to each of the 16 stimuli, resulting in $10,000$ simulated GCM data patterns and $10,000$ simulated MPM data patterns. For each simulated data pattern, we calculated the PSP fit of the GCM and MPM, using Equation (4.1), and we calculated a difference score: $\Delta\text{fit} = \text{PSP fit}_{\text{GCM}} - \text{PSP fit}_{\text{MPM}}$. A negative value of this difference score indicates a preference for the GCM, whereas a positive value indicates a preference for the MPM.

Next, we generated frequency distributions of the Δfit values for the patterns simulated from the GCM on the one hand and for the patterns simulated from the MPM on the other hand. Figure 4.6 depicts the Δfit distributions of the GCM patterns (in dark gray) and the MPM patterns

(in white) in each category structure.

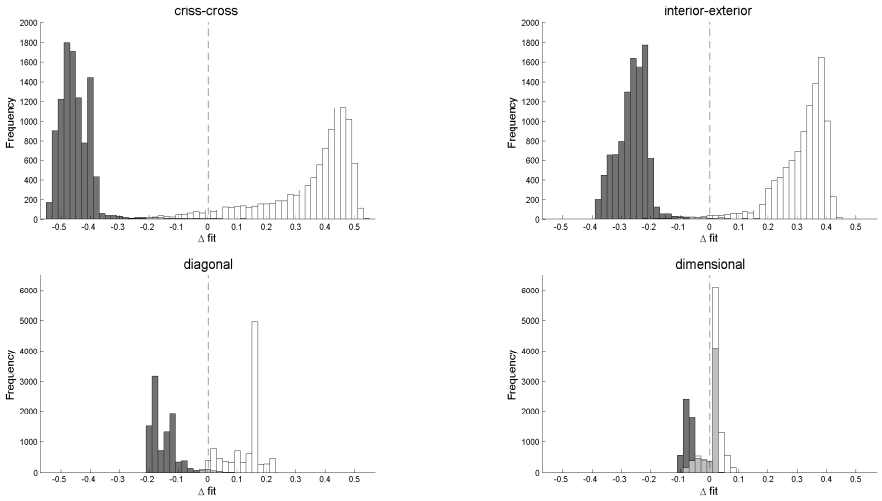


Figure 4.6: Δfit *distributions in the four different category structures, with* $\Delta\text{fit} = \text{fit}_{\text{GCM}} - \text{fit}_{\text{MPM}}$. *The dashed line indicates* $\Delta\text{fit} = 0$. *A negative value of* Δfit *indicates a preference for the GCM, whereas a positive value indicates a preference for the MPM. In each panel, the most left (dark gray) distribution reflects the* Δfit *distribution for the GCM patterns and the most right (white) distribution reflects the* Δfit *distribution for the MPM patterns. Light gray areas reflect overlapping parts of the GCM and MPM distributions.*

In the criss-cross structure, the GCM and MPM distribution lie for the greatest part at the left and at the right of $\Delta\text{fit} = 0$ (indicated by the dashed line), respectively, meaning that the majority of the simulated data patterns were correctly assigned to the model they were generated from. This observation is confirmed by the very high recovery rates, which indicate the proportion of simulated data sets that were assigned to the data-generating model (see Table 4.2). A small part of the MPM distribution is located at negative Δfit values, so there is a slight bias towards choosing the GCM in this structure.

A very similar picture, and similarly high recovery rates, emerge in the

Table 4.2: *Model recovery rates based on PSP fit for the GCM and the MPM in the four category structures from Nosofsky (1986).*

	Recovery rate	
Structure	GCM	MPM
Criss-cross	.9967	.9526
Interior-exterior	.9949	.9863
Diagonal	.9762	.9794
Dimensional	.5781	.8291

interior-exterior and diagonal structures. Again, the GCM and MPM distribution hardly cross $\Delta\text{fit} = 0$, so most of the data patterns were correctly assigned to the model they were generated from, based on their PSP fit.

Finally, in the dimensional structure, a large part of the GCM distribution is located at positive Δfit values (i.e., the light gray area to the right of the dashed line), and a smaller part of the MPM distribution is located at negative Δfit values (i.e., the light gray area to the left of the dashed line). This overlap is reflected in rather poor recovery rates, as shown in Table 4.2. These observations are consistent with our expectations based on the earlier observed indistinguishability of the models in this structure, and imply, again, that selecting between the GCM and the MPM would have been unwise in this structure, since it is impossible to know whether a GCM or MPM pattern is underlying most of the Δfit values.

Overall, for those structures where the PSP output in Figures 4.2 to 4.4 suggested that models are distinguishable, PSP based model selection showed excellent recovery. In the one condition where a lack of distinguishability gave an early warning that model selection was elusive, model selection, unsurprisingly, turned out to be very difficult.

Discussion

In this chapter, we proposed a way of quantitatively summarizing the output of a PSP analysis to assess a model's PSP fit, allowing for a straightforward decision rule for PSP based model selection. With an application in four category learning conditions, we demonstrated PSP based model selection between the GCM and the MPM. In three of the structures, PSP fit selected the GCM as the model with the best qualitative fit. In one structure, we refrained from selecting between the two models because an inspection of their largest data patterns suggested poor model distinguishability. The performance of PSP fit as a model selection method was evaluated with an extensive model recovery study. In those cases where a visual inspection of the PSP output suggested that models were distinguishable, PSP fit showed excellent recovery rates.

PSP based model selection differs from more traditional model selection methods (e.g., RMSE, AIC, BIC, Bayes factor, maximum likelihood, MDL) in the sense that it focuses on qualitative, instead of quantitative, data and model behavior, such as the ordinal relations between conditions. Thus, considering the PSP fit to select among models is useful for those researchers whose main interest is in a model's qualitative adequacy (e.g., McFall & Townsend, 1998; Pitt et al., 2003; Vanpaemel & Lee, 2012a; Wills & Pothos, 2012). Importantly, PSP based model selection is not intended to replace traditional model selection methods that focus on the quantitative properties of the data. Rather, it could be most fruitfully seen as complementary to these methods.

PSP fit assesses a model's scaled weighted average fit. By considering the average fit, model complexity is taken into account. PSP fit penalizes a model for generating data patterns that are dissimilar from the empirical pattern, as it increases with larger distances between these patterns. How-

ever, through V_i, the weight of a data pattern's distance depends on the representativeness of the pattern to the model. Generating data patterns very distant from the empirical trend is not that damaging for a model, provided that these patterns occupy a small part of the parameter space only. Likewise, generating data patterns similar to the observed one is only favorable for a model if these patterns are substantively representative for its behavior, that is, if these patterns are assigned a high volume.

By design, the application scope of PSP fit is the same as that of PSP. Thus, as PSP, PSP fit is currently not applicable to nonstationary models (see Pitt et al., 2006). Further, it is important to keep in mind that much like other model selection measures such as AIC or BIC, PSP fit is a measure of *relative* model fit. Thus, it can be used to compare fits of competing models, but it is mute about the absolute fit of the models under consideration.

While the recovery study indicated that PSP fit can perform well, we anticipate several future developments regarding PSP fit that are possible or even desired. First, this model evaluation approach is, in spirit, similar to the marginal likelihood from Bayesian model selection (Jeffreys, 1961; Kass & Raftery, 1995; Myung & Pitt, 1997) because the weighted average fit is considered. However, in contrast to Bayesian methods, PSP has a rather weak sensitivity to prior knowledge. In psychological models, priors can be used to express theoretical assumptions (e.g., Vanpaemel & Lee, 2012b). For example, an informative prior on the attention weight of the MPM and GCM can be used to express the attention-optimization hypothesis, which posits that learners tend to distribute their attention so as to optimize their performance (Nosofsky, 1986). Accordingly, when a prior expresses theory, model evaluation should be sensitive to the prior (Vanpaemel, 2010). PSP in its current form allows researchers to express assumptions about parameter values only by adapting the range of the parameter space. Future work could modify PSP so that a full prior distribution can be placed on the pa-

rameter space. The effect of an informative prior would be that the volumes of the regions are expanded or shrunk. Accordingly, data patterns would be weighted less or more heavily when they are generated by parameter values which are considered less or more plausible. Interestingly, PSP can provide useful information for determining priors. For example, if PSP output shows that the model generates implausible data patterns, the regions in the parameter space corresponding to these data patterns can be excluded or down-weighted in the prior (Lee & Vanpaemel, 2016).

Second, in its current implementation, PSP fit is insensitive to sample size, which may yield suboptimal selection behavior in some specific cases. In particular, if there is a very large amount of data available, sampling error diminishes and the data become almost free of noise. In this case, being able to predict the empirical data pattern should prevail in model selection. Stated differently, when sample size grows to infinity, a model that can predict the empirical data pattern exactly should be selected over a model that cannot generate this pattern, regardless of its complexity. This property is akin to the criterion of consistency in quantitative model selection procedures (where it is known that BIC is consistent, but AIC is not; see e.g. Vrieze, 2012). PSP fit may possibly penalize too heavily for model complexity when sample size is large. This observation has two consequences for the use of the current implementation of PSP fit. First, its use is most appropriate when sample size is small or moderate. Second, PSP fit is most suited for *explanatory*, rather than *predictive* modeling (see e.g., Shmueli, 2010). The purpose of explanatory modeling is to obtain the most accurate description of the (cognitive) process underlying the observed data. In this light, model complexity has to be penalized to overcome that a model is obtained that can fit many data patterns, without providing information about the underlying process. In predictive modeling, the goal is to predict new observations as accurately as possible, if necessary at the cost of theoretical

accuracy. In this framework, often large amounts of data are available, and complex models are required to generate accurate predictions. PSP fit may not be not suited for this latter type of modeling.

A third extension is related to the observation that, as with all model selection methods, conclusions based on PSP fit depend on the data that are observed. Whereas a given experiment can produce a certain data pattern, it might well be that other experiments generate completely different data patterns, resulting in different model selection outcomes. Wills and Pothos (2012) argued to move beyond model comparisons based on single data sets, and rather compare models across a broad range of data sets. Using PSP, the empirical representativeness of data patterns based on a literature review (i.e., the frequency with which they are observed in experiments) could be compared to the representativeness of the data patterns for the competing models (see Steingroever et al., 2013). A possible extension of PSP fit involves catering to this type of model comparison across a broad range of empirical data patterns.

Finally, an important advantage of PSP based model selection is that model distinguishability, which is important for any model selection method, can be assessed without recourse to additional extensive simulation studies. The same PSP output that lies at the basis of the PSP fit can serve as a proxy for distinguishability. Thus, *before* data collection and model selection, a researcher can run PSP on the models under consideration and examine whether their data patterns are distinctive enough. When the data patterns are very similar, the researcher is warned upfront that it will be difficult to tell the competing models apart, and that it is not worth the effort to collect empirical data. In this chapter, we relied on a visual representation of the most representative data patterns of the competing models to examine model distinguishability. The development of a formal way of assessing distinguishability that moves beyond visual inspection may be an interesting

direction for future research.

In conclusion, we hope that the current version of PSP fit is a useful first step in using the rich information provided by PSP for selecting between models based on qualitative aspects of models and data.

Appendices

Appendix A: Calculation of the PSP fit for the hypothetical models and data

In this section, we demonstrate how to compute the scaled average weighted PSP fits (Equation 4.1) of model 1 and model 2 to the empirical data pattern (3, 5, 2, 4, 1) under the Spearman's footrule distance, given by $d(DP_1, DP_2) = \sum_{j=1}^{m} |DP_{1j} - DP_{2j}|$, with DP_{1j} and DP_{2j} the rank position of element j in the first and second data pattern respectively. Like Figure 4.1, Figure 4.7 shows the models' parameter spaces partitioned into the regions that correspond to qualitatively different data patterns. In every region, the distance between the corresponding data pattern and the empirical pattern is depicted, multiplied with the proportion of the volume of the region. For example, the distance between the largest data pattern of model 1, (5, 1, 2, 3, 4), and the empirical data pattern is $|3-5| + |5-1| + |2-2| + |4-3| + |1-4| = 10$. Multiplying this distance with the volume of the corresponding region (i.e., 33%) results in a weighted distance of 3.3. Summing all the weighted distances of model 1 yields $3.3 + 0.4 + 2.4 + 1.7 + 1.2 + 0.0 + 0.66 = 9.66$. Under Spearman's footrule distance, the maximum distance between two data patterns with an odd number of elements is $\frac{(m+1)(m-1)}{2}$, where m is the number of elements, so in this example $d(DP^{\text{Emp}}, DP^{\text{Max}}) = 12$. Dividing the weighted distance by this maximal distance gives a PSP fit of 0.81 for model 1. A similar calculation results in a PSP fit of 0.17 for model 2.

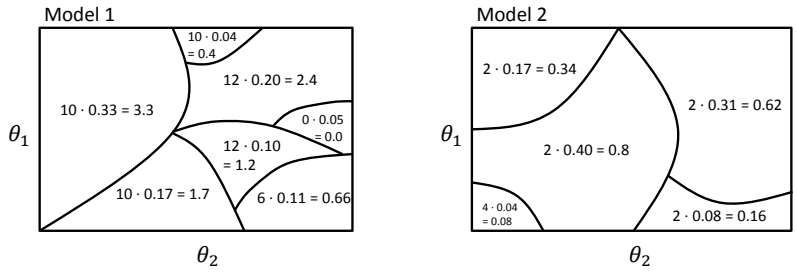


Figure 4.7: *The weighted distances between the fictional empirical data pattern (3, 5, 2, 4, 1) and the data patterns of hypothetical models, model 1 and model 2, as depicted in Figure 4.1.*

Appendix B: Formal details of the GCM and MPM

According to the Generalized Context Model (GCM; Nosofsky, 1986), classification decisions are based on the similarity of the stimulus to the relevant categories. As categories are assumed to be represented by individual exemplars, every exemplar in that category is considered in determining the similarity to a category. When there are two categories, A and B, the probability that stimulus i is classified into category A is formalized as:

$$P(A \mid i) = \frac{\sum_{j \in A} sim(i,j)}{\sum_{j \in A} sim(i,j) + \sum_{j \in B} sim(i,j)}, \tag{4.2}$$

where $sim(i,j)$ indicates the similarity between stimulus i and exemplar j, which is given by

$$sim(i,j) = \exp\left(-c\left[\sum_{m=1}^{M} w_m \mid x_{im} - x_{jm} \mid^r\right]^{\frac{1}{r}}\right). \tag{4.3}$$

In this equation, c is a sensitivity parameter, M is the number of stimulus dimensions, w_m is an attention-weight parameter, x_{im} and x_{jm} denote the coordinates of stimulus i and exemplar j on dimension m, respectively, and r is the metric. Most often, r is not considered a free parameter but is assumed

to depend on the type of dimensions that compose the stimuli. Generally, the city block metric ($r = 1$) is used when stimuli vary on separable dimensions, and the Euclidean metric ($r = 2$) is used for stimuli varying on integral dimensions. In the applications in this chapter, the Euclidean metric was used.

The Multiplicative Prototype Model (MPM; Minda & Smith, 2011; Nosofsky, 1987; Reed, 1972) differs from the GCM in one crucial assumption: A category is assumed to be represented by one abstract summary item, which means that, in determining the similarity of a stimulus to a category, only this item is considered. This summary item, called the prototype, is defined as the average of all individual exemplars. The coordinates of the category A prototype are the averaged coordinates of all of the n_A members of category A:

$$x_{Am} = \frac{1}{n_A} \sum_{j \in A} x_{jm}, \tag{4.4}$$

and likewise for category B. According to the MPM, the probability that stimulus i is classified into category A is formalized as:

$$P\left(A \mid i\right) = \frac{sim(i, A)}{sim(i, A) + sim(i, B)}, \tag{4.5}$$

where $sim(i, A)$ denotes the similarity between stimulus i and the category A prototype, given by

$$sim(i, A) = \exp\left(-c \left[\sum_{m=1}^{M} w_m \mid x_{im} - x_{Am} \mid^r\right]^{\frac{1}{r}}\right). \tag{4.6}$$

Appendix C: Detailed PSP procedure and output

For each of the four category structures used in Nosofsky (1986), we ran the PSP algorithm on both the GCM and MPM. The range of the parameter space of both models was $[0, 1]$ for the attention parameter w, and $[0, 20]$ for the sensitivity parameter c, based on what is typically observed in experiments. A data pattern was defined as the ranking of the category A response percentages (rounded to the nearest integer) for the 16 stimuli. All stimuli with equal rankings (i.e., same category A response percentages) were assigned a midrank value (i.e., rank position equal to the average of all the tied positions). As suggested by Pitt et al. (2006), we ran PSP five times for each model and structure to ascertain consistency. The PSP output comprised of all data patterns generated across the runs, with the volume of each data pattern determined by calculating the average of its five assigned volumes (if the data pattern had not been generated in a certain run, it was assigned a volume of 0%). This resulted in 8426, 14,848, 3902, and 6702 data patterns for the GCM, and 2142, 4444, 6222, and 5023 data patterns for the MPM in the criss-cross, interior-exterior, diagonal, and dimensional structure, respectively. The data patterns occupying at least one percent of the parameter space of the GCM and MPM can be found in the supplementary material, available on `https://osf.io/qe5kc`.

Appendix D: Recovery results sensitivity analysis

This section contains the results of the sensitivity analysis to evaluate the robustness of the PSP fits in the application example. Table 4.3 shows the PSP fits of the GCM and MPM in the four category learning structures obtained across the different data pattern definitions, considered stimuli and parameter ranges. The PSP fits between brackets reflect fits of models that appeared to be indistinguishable based on a visual inspection of their largest data patterns (the figures showing these data patterns in the diagonal structure can be found in the supplementary material, available on `https://osf.io/qe5kc`).

Table 4.3: *PSP fits for the GCM and the MPM in the four category learning structures from Nosofsky (1986) across two different data pattern definitions (ranking and nominal), two sets of stimuli (all stimuli and the unassigned stimuli only), and three different parameter ranges of the sensitivity parameter c ([0, 20],[0, 10] and [0, 5]). PSP fits between brackets reflect fits of indistinguishable models.*

		Ranking pattern				Nominal pattern			
		All stimuli		Unassigned stimuli		All stimuli		Unassigned stimuli	
Range c	Structure	GCM	MPM	GCM	MPM	GCM	MPM	GCM	MPM
$[0, 20]$	Criss-cross	0.17	0.59	0.29	0.61	0.01	0.50	0.01	0.50
	Interior-exterior	0.36	0.65	0.42	0.58	0.22	0.56	0.28	0.50
	Diagonal	0.13	0.30	(0.11)	(0.19)	0.05	0.20	(0.09)	(0.15)
	Dimensional	(0.19)	(0.26)	(0.21)	(0.27)	(0.002)	(0.05)	(0.002)	(0.06)
$[0, 10]$	Criss-cross	0.15	0.59	0.31	0.60	0.03	0.50	0.02	0.50
	Interior-exterior	0.36	0.65	0.42	0.58	0.24	0.56	0.30	0.50
	Diagonal	0.11	0.26	(0.10)	(0.15)	0.06	0.20	(0.08)	(0.15)
	Dimensional	(0.17)	(0.24)	(0.17)	(0.24)	(0.004)	(0.05)	(0.004)	(0.07)
$[0, 5]$	Criss-cross	0.14	0.57	0.33	0.59	0.04	0.50	0.03	0.50
	Interior-exterior	0.33	0.64	0.40	0.57	0.27	0.56	0.34	0.50
	Diagonal	0.09	0.23	(0.07)	(0.10)	0.08	0.20	(0.08)	(0.15)
	Dimensional	(0.15)	(0.20)	(0.14)	(0.20)	(0.01)	(0.05)	(0.007)	(0.06)

General Discussion

With this dissertation, we want to make a contribution to the larger movement in psychological research to critically scrutinize and improve the current research practices. More specifically, we focused on three commonly made advices for improving research practices. In the first chapter, we performed a replication study, implementing the most emphasized recommendation. In the second chapter, we elaborated on the advice to improve transparency by raising awareness of the necessity of an increased transparency about arbitrary choices in data processing. We demonstrated the importance of reporting the robustness of the results across these arbitrary choices by use of a multiverse analysis. In the last two chapters, we focused on the advice to adopt a Bayesian, rather than frequentist, statistical framework. In Chapter 3, two Bayesian model selection methods were compared: The commonly used Bayes factor versus the recently introduced Prior Information Criterion (PIC). The results revealed that the PIC, in contrast to the BF, can produce puzzling and undesirable outcomes. This chapter served as a cautionary tale: Although Bayesian testing is often recommended for better research practices, one should be critical about which specific Bayesian method to use. Finally, in Chapter 4, we extended the core idea of the Bayes factor – considering average fit rather than best fit – to qualitative data. We started from parameter space partitioning to propose a measure for qualitative model fit, which can be used for model selection when the interest is in qualitative data patterns.

In this final chapter, we discuss some additional reflections, and current, unpublished work on the good research practices that were addressed in this dissertation.

Replication

Replication has received a lot of attention in the light of the crisis of confidence in psychology. Many researchers have emphasized the importance of replication studies (e.g., Asendorpf et al., 2013; Finkel, Eastwick, & Reis, 2015; Pashler & Wagenmakers, 2012; Roediger, 2012), journals have shown an increased appreciation for replication attempts (e.g., Grahe, 2014; Lindsay, 2015; Nosek & Lakens, 2016; Zwaan & Zeelenberg, 2013), and several large-scale replication initiatives have emerged. Our replication study of the crowd within effect in Chapter 1 was also part of such a large-scale replication project, known as the Reproducibility Project: Psychology (RPP; Open Science Collaboration, 2015).

The RPP was a collaborative effort to evaluate the reproducibility of published findings in psychology. In total, 100 replication studies were performed by different groups of researchers, including our crowd within study. Whereas we successfully replicated the original finding, this was not the case for all studies: Only 36% of the replications had significant results (in contrast to 97% of significant results in the original studies) and 47% of the original effect sizes were in the 95% confidence interval of the replication effect size. The publication of these findings had a large impact in the research field, as it fostered awareness of the need for improved research practices in the whole psychological scientific community.

Other large-scale replication efforts in which we are currently collaborating are the Many Labs 5 project (Ebersole et al., 2017) and the Action-sentence Compatibility Effect Preregistered Replication (Kaschak et al., 2017).

The goal of Many Labs 5 is to evaluate whether formal peer review of replication protocols can improve reproducibility. The project takes the 11 replications from the RPP that were labeled as "not endorsed" as a starting point. In the RPP, all replication teams contacted the authors from the original studies and asked them for feedback on a draft of the replication design before data collection. Whereas most of the original authors endorsed these replication protocols, in 11 cases, concerns were expressed about factors that could interfere replication such as a different sample or changes to the design. In the Many Labs 5 project, these 11 non-endorsed replication protocols are revised by the original authors and/or other domain experts until they are endorsed. Afterwards, new replications, both with the non-endorsed RPP protocols and the revised protocols, are conducted by different replications teams, where (if possible) each team randomly assigns participants to either the endorsed or non-endorsed protocol. A comparison of the replication results obtained from these two types of protocols will shed light on the question as to whether formal peer review on replication protocols can enhance reproducibility.

Our contribution to the Many Labs 5 project is a collaboration on the replication of Study 6 by Risen and Gilovich (2008), which examines the belief that "tempting fate" is punished with ironic bad outcomes. In this study, participants read a scenario which asks them to imagine themselves in a large lecture and they either read that they have done the reading for the class or that they have not done the reading. Then, they are asked how likely they believe it is that they would be called on by the professor to answer a question. Half of the participants are placed under a cognitive load during reading and answering by making them count backwards by 3s. The results from the original study showed that participants who imagined that they had not done the reading believed that it was more likely that they would be called on in class than participants who imagined that they had

done the reading, and this effect was more pronounced when participants responded under cognitive load. The authors interpreted these findings as evidence for their hypothesis that cognitive System 2 processes (which are are preoccupied by a cognitive load) can suppress the irrational belief in tempting fate, originating from heuristic System 1 processes. However, the (non-endorsed) RPP replication attempt failed to replicate these original findings. Concerns that were raised by the original authors about the RPP replication protocol were related to the sampling frame (undergraduates in the original study vs. Mechanical Turk participants in the replication study) and questionnaire administration (on paper in the original study vs. online in the replication study). They argued that the scenario in the questionnaire would be more relatable to undergraduates than to a non-student sample, and that an online environment could interfere adequate control of the cognitive load. The revised and endorsed protocol therefore differs from the non-endorsed protocol in the following ways: The sampling frame consists of undergraduate students instead of MTurkers, the procedure is conducted in person instead of online, and manipulation checks are included. Unlike in other studies in the Many Labs 5 project, random assignment of participants to endorsed and non-endorsed protocols within replication teams was not possible here, because of the different sampling frames in the two protocols. Therefore, different replication teams conducted the replication study with either the endorsed protocol (which was the case for our replication team) or the non-endorsed protocol. At the time of writing, all data are collected and under analysis.

The Action-sentence Compatibility Effect Preregistered Replication is a multi-lab replication study to examine the reproducibility of the Action-sentence Compatibility Effect (ACE; Glenberg & Kaschak, 2002), which is a motor compatibility effect that indicates a relationship between language comprehension and motor processing. More specifically, the effect is

demonstrated by letting participants judge whether sentences describing directional actions (e.g., "Meghan handed you the book" describes an action *toward* you, whereas "You handed Meghan the book" describes an action *away* from you) are sensible or not by pressing either a button near the body (a *toward* response) or a button farther from the body (an *away* response). According to the ACE, sensibility judgments are made faster if the direction of the sentence action matches the direction of the response action. Whereas this effect has been important for theory building in language comprehension, its reproducibility was recently questioned by Papesh (2015). However, the lead researchers from the ACE Preregistered Replication disagree with the strength of evidence presented against ACE. Therefore, they established a large-scale replication effort of the original ACE study with the collaboration of different replication teams, in order to resolve questions concerning the reproducibility of ACE.

Naturally, this emerging emphasis on replication studies in psychology comes with new challenges. A current issue is the establishment of clear standards for evaluating replication results. In Chapter 1, we evaluated our replication results against two replication standards. Besides the traditional approach of regarding a replication as successful or failed if a significant or non-significant result is obtained, we also used the "detectability approach", as introduced by Simonsohn (2015). This approach evaluates replication results by assessing whether they are consistent with an effect size big enough so that it could have been detected in the original study. On the same line of reasoning, Verhagen and Wagenmakers (2014) proposed a Bayesian replication test to quantify the extent to which a replication attempt has failed or succeeded by evaluating whether the effect from a replication study is comparable to the effect that was found in the original study. Others have advocated to evaluate replication results by conducting a meta-analysis, in which the original study result is combined with the replication result(s) to

calculate an average weighted effect size (e.g., Asendorpf et al., 2013; Rouder & Morey, 2011; Valentine et al., 2011). Clearly, there are different ways to examine replication results, and it is probably useful to consider multiple perspectives. For example, in the RPP, the replication results were evaluated according to five criteria, including p-values, effect sizes, meta-analysis and subjective assessments of replication. However, it would be useful if future work would establish more clear guidelines concerning evaluation standards for replication results.

Another issue is the distinction between direct and conceptual replications. In Chapter 1, we performed a direct replication study, meaning that we tried to exactly repeat the original crowd within study as close as possible, using the same methodology. Successful exact replications can confirm the replicability of findings. On the other hand, conceptual replication studies aim to examine the generalizability of an effect by conducting the experiment using a different methodology, context or population. In the context of the crisis of confidence, most researchers have stressed the importance of exact replication studies, since failed conceptual replication attempts leave open too many interpretations for the reason of failure (e.g., Nosek, Spies, & Motyl, 2012; Simons, 2014). However, both replicability and generalizability are necessary to evaluate the thrustworthiness of findings (LeBel, Vanpaemel, McCarthy, Earp, & Elson, 2017).

In a recent pilot project on which I collaborated, we worked on a method to examine the generalizability of an effect across different experimental contexts, using a single study. More specifically, the approach focuses on the generalizability of results across different design decisions, such as the selection of stimuli, the wording of instructions, the measurement of the dependent variable, and so on. Different decisions, resulting in different "experimental contexts", can lead to different, heterogeneous results. A well-known and often recommended method to deal with this heterogene-

ity is meta-analysis, which combines the results from different studies to estimate the between-study variability and the generalizability of the effect across studies. Based on this idea, we propose a more systematic and efficient way to evaluate the generalizability of an effect across different experimental contexts, namely a *meta-study*. In a meta-study, a researcher lists all reasonable experimental choices, resulting in a population of experimental contexts, and then studies the effect in a random sample from this population, with only a few participants assigned to each context. Thus, a meta-study comprises one large study, consisting of a sample of small studies with different experimental contexts. With the use of a hierarchical analysis, this approach will allow researchers to assess and report the robustness and generalizability of an effect across different experimental designs, based on a single study.

Transparency

Increasing transparency is another frequently suggested solution for restoring confidence in psychological research findings. This topic covers a diversity of recommendations, related to different aspects of the research process, such as preregistration, reporting the sample size plan, reporting all measures, open materials, open data, and so on (e.g., Asendorpf et al., 2013; Ioannidis, 2014; Nosek et al., 2015; Rouder, 2016; Simmons et al., 2012; Wagenmakers et al., 2012; Wicherts et al., 2011). In our crowd within replication study (Chapter 1), we fully committed to transparency practices by preregistering the study to the level of the analysis code, and making all study details publicly available at the Open Science Framework, including the experimental materials, the raw and processed data, and the analysis code.

In Chapter 2, we highlighted the importance of being transparent about one specific aspect of research, namely arbitrary choices in data process-

ing. We argued that researchers should move away from a single data set analysis, and disclose the full multiverse of results across different arbitrary choices in data processing. Recently, Credé and Phillips (2017) performed a multiverse analysis to examine the robustness of the "power pose" effect (i.e., the effect that body postures associated with power result in increased levels of testosterone and risky decision-making behavior) across alternative data analytic specifications. They considered three data analytic decisions, including arbitrary choices in data processing (identification of outliers), as well as in model choice (choice of dependent variable, use of control variables). The multiverse of results revealed that the power pose effect is highly sensitive to the specific combination of these choices.

Silberzahn et al. (2017) suggested another way for dealing with the issue of many potential outcomes: They evaluated the impact of the diversity in analytic choices by crowdsouring data analysis. Different research teams analyzed the same data set to answer the same research question concerning the relation between skin tone and red cards in soccer. Each team subjectively decided how they would analyze the data set in order to answer the same hypothesis. The results showed not only a high variability in analytical approaches – all based on reasonable decisions – across teams, but also an accompanying high variability in estimated effect sizes.

The movement towards an increased transparency in research obviously also comes with many challenges. For example, whereas there is a growing appreciation for preregistration, clear preregistration standards are lacking. Wagenmakers et al. (2012, p. 635) suggested that "before a single participant is tested, the researcher submits ... a document that details what dependent variables will be collected and how the data will be analyzed (i.e., which hypotheses are of interest, which statistical tests will be used, and which outlier criteria or data transformations will be applied)". Indeed, this captures the core a what a preregistration document should contain.

However, a more detailed checklist would be useful. For example, Wicherts et al. (2016) presented a checklist of 34 degrees of freedom in the planning, executing, analyzing and reporting of psychological studies, which can be useful for assessing the quality of preregistrations. In a recent project on which I collaborated, a preliminary version of a detailed preregistration checklist has been created. The purpose of this checklist is to make preregistration more accessible to researchers by providing them clear guidelines, and to improve the quality of research by making researchers aware of the necessary components in a preregistration document.

Another unresolved issue is how to handle studies that deviated from the preregistration document. Preregistered studies are considered to be of higher quality, but not all studies adhere to the registered protocol (LeBel et al., 2017). For example, the COMPare team (see `compare-trials.org`) recently compared all clinical trial reports that were published in the top five medical journals with their registered protocols. They found that, on average, only 58.2% of the prespecified outcomes were reported and 5.3 new outcomes were silently added in each trial. It may be interesting for future work to establish guidelines for evaluating the level of adherence to the preregistered document and how to handle different types of non-adherence.

Bayesian Framework

Another frequently mentioned recommendation for improving research practices is the adoption of a Bayesian, rather than frequentist framework (e.g., Dienes & McLatchie, 2017; Rouder et al., 2009; Wagenmakers et al., 2011; Wagenmakers, Morey, & Lee, 2016). Especially the use of Bayes factors is highlighted in this class of recommendations, as they are the Bayesian answer to hypothesis testing, which still is the workhorse approach for most psychologists. Advantages of Bayesian methods include, among others, insensitivity to the stopping rule that is used for data collection, the ability

to handle small data sets, and the ability to provide evidence for both the alternative and the null hypothesis.

Convinced of the conceptual and practical benefits of Bayesian inference compared to traditional frequentist inference, recent efforts have been made to overcome computational obstacles for calculating Bayes factors. The development of user-friendly software such as JASP (JASP Team, 2017) and the `BayesFactor` package in R (Morey & Rouder, 2014) now allows researchers to easily compute Bayes factors for standard methods, such as *t*-tests, ANOVA and regression.

Another obstacle to the adoption of Bayesian methods is that they require the specification of a prior distribution. Not only is it often difficult to find a suitable distribution to capture prior knowledge, some authors also find priors problematic because they further increase researcher degrees of freedom, creating extra opportunities for questionable research practices (e.g., Simmons et al., 2011). However, as we already noted in Chapters 3 and 4, we find the implementation of priors very useful, as they can be used to express theoretical assumptions. Moreover, just as other with subjective data analytic choices, this problem can be dealt with easily by checking the robustness of the results across different reasonable prior specifications. Whereas most user-friendly tools for calculating Bayes factors make use of default priors, it would be interesting for future research to expand this to informative priors. A recent example can be found in Gronau, Ly, and Wagenmakers (2017), who introduced an extension to the default Bayesian *t*-test (Jeffreys, 1948; Rouder et al., 2009), allowing researchers to assign an informed prior distribution to effect sizes. In contrast to the zero-centered Cauchy prior, which is the default in the `BayesFactor` package, expert knowledge about the effect size can be captured by this informed prior, centered away from zero.

Gronau et al. (2017) demonstrated this informed Bayesian *t*-test with

three application examples, one of which was a reanalysis of the crowd within data that we collected in our replication study in Chapter 1. More specifically, this reanalysis focused on the comparison of the first guess to the average guess in the delayed condition, for which we found that the error of the average was smaller than the error of the first guess, $t(139) = 4.02, p < .001; \text{BF} = .007$ (BF in favor of the null hypothesis of no effect; see Table 1.5). The reanalysis of these data was "informed" by the results from the original study from Vul and Pashler (2008). That is, first a posterior distribution for the effect size was generated based on the results from Vul and Pashler (2008). This posterior was then used as an informed prior distribution for our replication data; A Bayes factor was calculated, comparing the null hypothesis of a zero effect size with the alternative hypothesis assigning this informed prior to the effect size. The obtained Bayes factor was BF = 901.5 in favor of the informed alternative hypothesis. In comparison, the default Bayesian t-test (which compares the null hypothesis to the zero-centered Cauchy prior) yielded a BF = 170.2 (when the scale parameter for the Cauchy is set to $1/\sqrt{2}$) or BF = 132.8 (when the scale parameter for the Cauchy is set to 1) in favor of the alternative. Thus, the informed Bayesian t-test provides more evidence for the alternative than the default test, which makes logic sense since the former takes into account the expectations based on the original study, and the replication results are in line with those expectations.

Bayesian vs frequentist: Does it matter?

Despite the growing sense of the advantages of using a Bayesian, rather than a frequentist framework, a question which remains unanswered is whether this would make a difference in practice: If the statistical workhorse in psychology would change from p-values to Bayes factors, would there be a change in qualitative results? In a current project, we aim to address how

psychology would look like from a Bayesian perspective by evaluating the extent to which conclusions based on Bayes factors differ from to those based on p-values.

Wetzels et al. (2011) performed an empirical comparison between Bayes factors and p-values using 855 published t-tests in psychology. For each of the 855 empirical findings, they calculated a default Bayes factor and compared it with the corresponding p-value. Whereas both measures greatly covaried, the main difference was that p-values tended to provide more support in favor of the alternative hypothesis compared to Bayes factors. On a practical level, many of the published significant results would have been considered "worth no more than a bare mention" if Bayes factors instead of p-values would have been used for inference. Whereas Wetzels et al. (2011)'s empirical comparison was restricted to t-tests, the current project covers a much broader comparison between Bayes factors and p-values, applied to a wide variety of research designs.

In order to compare the results based on Bayes factors versus those based on p-values, we will reanalyze data from studies that were published in the 2012 issues of four different APA journals: *Emotion*, *Experimental and Clinical Psychopharmacology*, *Journal of Abnormal Psychology* and *Psychology and Aging*. Vanpaemel, Vermorgen, Deriemaecker, and Storms (2015) collected the data by emailing the first authors of all of the concerning articles with an explanation of the purpose of our project and a request to send the raw data from their study. In total, Vanpaemel et al. (2015) asked approximately 400 data sets and received approximately 150 data sets. This response rate is disconcertingly low, but higher than the sharing rate that was found by Wicherts, Borsboom, Kats, and Molenaar (2006) a few years earlier.

In each data set, we will focus on the main results only, based on what is written in, for example, the abstract or the conclusion of the corresponding

article. Further, we will only focus on results based on analyses for which the equivalent Bayes factor calculation is implemented in the `BayesFactor` package, as we are interested in results from Bayes factors that could have been easily obtained by the original authors.

In a first stage, we are trying to replicate the original main results by performing the same frequentist analyses as reported in the article. Unexpectedly, this turned out to be not straightforward. Obstacles that we have encountered in this process are, for example, difficulties to find out which statistical analyses were performed, which variables were selected, which participants were selected, and not enough data available to replicate the analyses. In the first two journals that we already considered (i.e., *Journal of Abnormal Psychology* and *Psychology and Aging*), we encountered reproducibility issues – ranging from obtaining different degrees of freedom to not knowing which variables to select – in approximately two-thirds of the articles. We were surprised to encounter these kind of problems for such a large number of articles, and we believe this flagrantly confirms the need for an increased transparency, such as the sharing of replication files (Dafoe, 2014).

The next step is to contact the original authors of the articles with reproducibility issues and ask them for clarification. Once we have managed to overcome the reproducibility hurdles and replicated (or corrected) all main findings, we will calculate the equivalent Bayes factor, using the `BayesFactor` package. Across all results, we will compare the conclusions based on the p-values with those based on the Bayes factors.

Final Note

The crisis of confidence in psychology has triggered many researchers to propose recommendations for better research practices. These recommendations cover a wide variety of topics across the whole research cycle. In this

dissertation, we addressed some of these topics, but others remained largely undiscussed. For example, it also has been argued that organizations such as granting agencies, tenure committees and the educational system should place more emphasis on responsible conduct of research, or that journals should change their publication policies. Clearly, there is no single solution to overcome the crisis of confidence, but hopefully this dissertation can make a valuable contribution towards better research practices in psychology.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & B. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267–281). Budapest: Academiai Kiado.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., ... others (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*, 372–400.

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Banker, S., & McCoy, J. (2013). *Replication attempt crowd within.* (unpublished raw data)

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*, 345–370.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., ... Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Central Intelligence Agency. (2013). *The world factbook.* Retrieved June 7, 2013, from `https://www.cia.gov/library/publications/the-world-factbook/`

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610.

Claeskens, G., & Hjort, N. L. (2008). *Model selection and model averaging* (Vol. 330). Cambridge University Press Cambridge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, New Jersey: Lawrence Erlbaum.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *Handbook of research synthesis and meta-analysis.* Russell Sage Foundation.

Credé, M., & Phillips, L. A. (2017). Revisiting the power pose effect: How robust are the results reported by Carney, Cuddy, and Yap (2010) to data analytic decisions? *Social Psychological and Personality Science*, 1948550617714584.

Cumming, G. (2013). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis.* Routledge.

Dafoe, A. (2014). Science deserves better: The imperative to share complete replication files. *PS: Political Science & Politics*, *47*, 60–66.

Dani, V., Madani, O., Pennock, D. M., Sanghai, S., & Galebach, B. (2012). An empirical comparison of algorithms for aggregating expert predictions. *arXiv preprint arXiv:1206.6814*.

De Groot, A. (1956/2014). The meaning of significance for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny

Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han LJ van der Maas]. *Acta Psychologica*, *148*, 188–194.

Diaconis, P., & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, 262–268.

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89.

Dienes, Z., & McLatchie, N. (2017). Four reasons to prefer Bayesian over orthodox statistical analyses. *Psychonomic Bulletin and Review*.

Doob, J. L. (1949). Application of the theory of martingales. *Le calcul des probabilites et ses applications*, *13*, 23–27.

Durante, K. M., & Arsena, A. R. (2015). Playing the field: The effect of fertility on women's desire for variety. *Journal of Consumer Research*, *41*, 1372–1391.

Durante, K. M., Arsena, A. R., & Griskevicius, V. (2014). Fertility can have different effects on single and nonsingle women: Reply to Harris and Mickes (2014). *Psychological Science*, *25*, 1150–1152.

Durante, K. M., Griskevicius, V., Cantú, S. M., & Simpson, J. A. (2014). Money, status, and the ovulatory cycle. *Journal of Marketing Research*, *51*, 27–39.

Durante, K. M., Griskevicius, V., Hill, S. E., Perilloux, C., & Li, N. P. (2011). Ovulation, female competition, and product choice: Hormonal influences on consumer behavior. *Journal of Consumer Research*, *37*, 921–934.

Durante, K. M., Griskevicius, V., Simpson, J. A., Cantú, S. M., & Li, N. P. (2012). Ovulation leads women to perceive sexy cads as good dads. *Journal of Personality and Social Psychology*, *103*, 292–305.

Durante, K. M., Rae, A., & Griskevicius, V. (2013). The fluctuating female vote: Politics, religion, and the ovulatory cycle. *Psychological Science*, *24*, 1007–1016.

Ebersole, C. R., Nosek, B. A., Kidwell, M., Buttrick, N., Baranski, E., Chartier, C. R., ... et al. (2017, Jan). *Many labs 5: Can conducting formal peer review in advance improve reproducibility?* Open Science Framework. Retrieved from `osf.io/7a6rd`

Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., & Vee, E. (2006). Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, *20*, 628–648.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Finkel, E. J., Eastwick, P. W., & Reis, H. T. (2015). Best research practices in psychology: Illustrating epistemological and pragmatic considerations with the case of relationship science. *Journal of Personality and Social Psychology*, *108*, 275–297.

Galton, F. (1907). Vox populi. *Nature*, *75*, 450–451.

Gangestad, S. W., Haselton, M. G., Welling, L. L., Gildersleeve, K., Pillsworth, E. G., Burriss, R. P., ... Puts, D. A. (2016). How valid are assessments of conception probability in ovulatory cycle research? Evaluations, recommendations, and theoretical implications. *Evolution and Human Behavior*, *37*, 85–96.

Geisser, S. (1993). Perturbation analysis. In *Predictive Inference: An introduction* (pp. 129–153). Springer.

Gelman, A., Hill, J., & Yajima, M. (2012). Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, *5*, 189–211.

Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist*, *102*, 460–465.

Ghosal, S., Ghosh, J. K., & van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, *28*, 500–531.

Gilks, W., Richardson, S., & Spiegelhalter, D. J. (Eds.). (1996). *Markov chain Monte Carlo in practice*. Suffolk, UK: Chapman and Hall.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review*, *9*, 558–565.

Grahe, J. E. (2014). *Announcing open science badges and reaching for the sky.* Taylor & Francis.

Gronau, Q., Ly, A., & Wagenmakers, E. (2017). Informed Bayesian t-tests. *arXiv preprint arXiv:1704.02479*.

Harris, C. R. (2013). Shifts in masculinity preferences across the menstrual cycle: Still not there. *Sex Roles*, *69*, 507–515.

Harris, C. R., Pashler, H., & Mickes, L. (2014). Elastic analysis procedures: An incurable (but preventable) problem in the fertility effect literature. Comment on Gildersleeve, Haselton, and Fales (2014). *Psychological Bulletin*, *140*, 1260–1264.

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

Hawkins, G. E., Brown, S. D., Steyvers, M., & Wagenmakers, E.-J. (2012). An optimal adjustment procedure to minimize experiment time in decisions with multiple alternatives. *Psychonomic Bulletin & Review*, *19*, 339–348.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231–237.

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*, 1068-1074.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS medicine*, *11*, e1001747.

Ioannidis, J. P., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*, 235–241.

JASP Team. (2017). *JASP (Version 0.8.1.1).* Computer software.

Jeffreys, H. (1948). *Theory of probability* (2nd ed.). Oxford: Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.

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

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*, 19313–19317.

Jones, M., & Love, B. (2011). Bayesian fundamentalism or enlightenment? On the explanatory status and theoretical contributions of Bayesian models of cognition. *Behavioral and Brain Sciences*, *34*, 169–188.

Kaschak, M., Zwaan, R. A., Glenberg, A., Morey, R. D., Ibanez, A., Gianelli, C., ... et al. (2017, Jun). *Action-sentence compatibility effect (ACE) pre-registered replication.* Open Science Framework. Retrieved from `osf.io/ynbwu`

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81-93.

Kendall, M. G. (1975). *Rank correlation methods* (4th ed.). London: Charles Griffin.

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*, 196–217.

Kleijn, B. J. K., & van der Vaart, A. W. (2006). Misspecification in infinite-dimensional Bayesian statistics. *Annals of Statistics*, *34*, 837–887.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22-44.

Kruschke, J. K. (2010). What to believe: Bayesian methods for data analysis. *Trends in Cognitive Sciences*, *14*, 293–300.

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299–312.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, *73*, 31–43.

Leamer, E. E. (1985). Sensitivity analyses would help. *The American Economic Review*, *75*, 308–313.

LeBel, E. P., Campbell, L., & Loving, T. J. (in press). Benefits of open and high-powered research outweigh costs. *Journal of Personality and Social Psychology*.

LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, *15*, 371–379.

LeBel, E. P., Vanpaemel, W., McCarthy, R. J., Earp, B. D., & Elson, M. (2017, Jun). *A unified framework to quantify the trustworthiness of empirical research.* Open Science Framework. Retrieved from `osf.io/afrzx`

Lee, M. D., & Vanpaemel, W. (2016). Determining informative priors for cognitive models. *Psychonomic Bulletin & Review*, 1–14.

Lindsay, D. S. (2015). *Replication in psychological science.* SAGE Publications.

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods*, *7*, 19–40.

MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth. *Nature*, *526*, 187–190.

Marden, J. I. (1995). *Analyzing and modeling rank data.* London: Chapman Hall.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McFall, R. M., & Townsend, J. T. (1998). Foundations of psychological assessment: Implications for cognitive assessment in clinical science. *Psychological Assessment*, *10*, 316–330.

Minda, J. P., & Smith, J. D. (2011). Prototype models of categorization: Basic formulation, predictions, and limitations. In E. Pothos & A. Wills (Eds.), *Formal Approaches in Categorization.* Cambridge, UK: Cambridge University Press.

Morey, R. D., Chambers, C. D., Etchells, P. J., Harris, C. R., Hoekstra, R., Lakens, D., ... others (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, *3*, 150547.

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*, 406–419.

Morey, R. D., & Rouder, J. N. (2014). Package BayesFactor.

Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data*

Analysis, *71*, 448–463.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences*, *23*, 299–325.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S., Breckler, S., ... others (2015). Promoting an open research culture: Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science*, *348*, 1422–1425.

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry*, *23*, 217–243.

Nosek, B. A., & Lakens, D. (2016). Registered reports: A method to increase the credibility of published reports.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.

Nosofsky, R. M. (1987). Attention and learning processes in the identification and categorization of integral stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 87–108.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

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

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science a crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530.

Patel, C. J., Burford, B., & Ioannidis, J. P. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of clinical epidemiology*, *68*, 1046–1058.

Pitt, M. A., Kim, W., & Myung, J. I. (2003). Flexibility versus generalizability in model selection. *Psychonomic Bulletin & Review*, *10*, 29–44.

Pitt, M. A., Kim, W., Navarro, D. J., & Myung, J. I. (2006). Global model analysis by parameter space partitioning. *Psychological Review*, *113*, 57–83.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Pitt, M. A., Myung, J. I., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Ramsey, F., & Schafer, D. (2012). *The statistical sleuth: A course in methods of data analysis* (3rd ed.). Stanford: Cengage Learning.

Rauhut, H., & Lorenz, J. (2011). The wisdom of crowds in one mind: How individuals can simulate the knowledge of diverse societies to reach better decisions. *Journal of Mathematical Psychology*, *55*, 191–197.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 393-407.

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology*, *95*, 293–307.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological review*, *107*, 358–367.

Roediger, H. (2012). Psychology's woes and a partial cure: The value of replication. *APS Observer*, *25*, 9.

Rouder, J. N. (2016). The what, why, and how of born-open data. *Behavior research methods*, *48*, 1062–1069.

Rouder, J. N., & Morey, R. D. (2011). A Bayes factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review*, *18*, 682–689.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Schervish, M. J. (1995). *Theory of statistics*. New York: Springer.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological methods*, *17*, 551–566.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *4*, 10–26.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*. (13, Whole No. 517)

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*, 289–310.

Sijtsma, K. (2016). Playing with data – or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika*, *81*, 1–15.

Silberzahn, R., Uhlmann, E. L., Martin, D., Aust, F., Awtrey, E. C., Bahník, Š., ... others (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). A 21 word solution.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80.

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, 0956797614567341.

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72-101.

Spies, J. R., Nosek, B. A., Miguez, S., Blohowiak, B. B., Cohn, M., Bartmess, E., ... Cohoon, J. (2012, May). *Open science framework.* Retrieved June 6, 2013, from `http://opensciencframework.org/`

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Corrigendum: Measuring the crowd within again: A pre-registered replication study. *Frontiers in psychology*, *6*.

Steingroever, H., Wetzels, R., & Wagenmakers, E.-J. (2013). A comparison of reinforcement learning models for the Iowa gambling task using parameter space partitioning. *The Journal of Problem Solving*, *5*, Article 2.

Steyvers, M., Lee, M., Miller, B., & Hemmer, P. (2009). The wisdom of crowds in the recollection of order information. In J. Lafferty &

C. Williams (Eds.), *Advances in neural information processing systems.* MIT Press.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71.

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations.* New York, NY: Doubleday.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Science*, *10*, 309–318.

Thornhill, R., & Gangestad, S. W. (1999). The scent of symmetry: A human sex pheromone that signals fitness? *Evolution and Human Behavior*, *20*, 175–201.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., ... Schinke, S. P. (2011). Replication in prevention science. *Prevention Science*, *12*, 103–117.

van de Schoot, R., Hoijtink, H., Romeijn, J.-W., & Brugman, D. (2012). A prior predictive loss function for the evaluation of inequality constrained hypotheses. *Journal of Mathematical Psychology*, *56*, 13–23.

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*, 491–498.

Vanpaemel, W. (2016). Prototypes, exemplars and the response scaling parameter: A Bayes factor perspective. *Journal of Mathematical Psychology*, *72*, 183–190.

Vanpaemel, W., & Lee, M. D. (2012a). The Bayesian evaluation of categorization models: Comment on Wills and Pothos (2012). *Psychological Bulletin*, *138*, 1253–1258.

Vanpaemel, W., & Lee, M. D. (2012b). Using priors to formalize theory: Optimal attention and the generalized context model. *Psychonomic*

Bulletin & Review, *19*, 1047–1056.

Vanpaemel, W., & Storms, G. (2010). Abstraction and model evaluation in category learning. *Behavior Research Methods*, *42*, 421–437.

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra: Psychology*, *1*, 1–5.

Verhagen, J., & Wagenmakers, E.-J. (2014). Bayesian tests to quantify the result of a replication attempt. *Journal of Experimental Psychology: General*, *143*, 1457–1475.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the Akaike information criterion (AIC) and the Bayesian information criterion (BIC). *Psychological Methods*, *17*, 228–243.

Vul, E. (n.d.). *Crowd within.* Retrieved June 7, 2013, from `http://www.edvul.com/crowdwithin.php`

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, *4*, 274–290.

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*, 645–647.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*, 169–176.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research.

Perspectives on Psychological Science, *7*, 632–638.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi. *Journal of Personality and Social Psychology*, *54*, 426–432.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology an empirical comparison using 855 t tests. *Perspectives on Psychological Science*, *6*, 291–298.

Wicherts, J. M. (2011). Psychology must learn a lesson from fraud case. *Nature*, *480*, 7.

Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS one*, *6*, e26828.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*, 726–728.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological Bulletin*, *138*, 102-125.

Yong, E. (2012). In the wake of high profile controversies, psychologists are facing up to problems with replication. *Nature*, *485*, 298–300.

Zwaan, R., & Zeelenberg, R. (2013). Replication attempts of important results in the study of cognition. *Frontiers in Cognition*.