Possible Solution to Publication Bias through Bayesian Statistics, including Proper Null Hypothesis Testing

Elly A. Konijn[1], Rens van der Schoot[2,3], Sonja Winter[2], & Christopher J. Ferguson[4]

[1] *VU University Amsterdam, Dept. of Communication Science, Media Psychology Program, and Netherlands Institute for Advanced Studies (NIAS), Wassenaar, The Netherlands*

[2] *Utrecht University, Dept. of Methods and Statistics, The Netherlands*

[3] *Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark, South Africa.*

[4] *Stetson University, Department of Psychology, USA*

4th REVISED Paper submitted to *Communication Methods and Measures - Special Issue on "Questionable Research and Publication Practices in Communication Science."*

Revision-4 submitted d.d. 21 July 2015

Special Issue Editors: Tilo Hartmann & Ivar Vermeulen

* Correspondence to: Elly A. Konijn, professor of Media Psychology, Dept. of Communication Science, VU University Amsterdam, Metropolitan Building, Rm. Z.4.14, 1081 HV Amsterdam, The Netherlands. Tel.: +31 205986839; E-mail address: e.a.konijn@vu.nl

Abstract

The present paper goes beyond discussing publication bias as currently prevalent in the social sciences, including communication science, by arguing that an important cause of publication bias is the way in which traditional frequentist statistics force binary decisions. We propose an alternative approach through Bayesian statistics. Bayesian statistics provide various degrees of support for any hypothesis allowing balanced decisions and proper null hypothesis testing, which may prevent publication bias. Moreover, to test a null hypothesis becomes increasingly relevant in mediated communication and virtual environments. To illustrate our arguments, we re-analyzed three datasets of previously published data (i.e., media violence effects; mediated communication; visuospatial abilities across genders). Results are discussed in view of possible interpretations, which are more open to a content-related argumentation of the level of support by means of Bayes Factors. Finally, we discuss potential pitfalls of a Bayesian approach such as BF-hacking if cut-off values are going to be introduced as in classical hypothesis testing: "God would love a Bayes Factor of 3.01 nearly as much as a BF of 2.99" (cf. Rosnow & Rosenthal, 1989). Especially when BF values are small, replication studies and Bayesian updating are still necessary to draw conclusions.

[word count: 195]

Possible Solution to Publication Bias through Bayesian Statistics, including Proper Null Hypothesis Testing

"Our unfortunate historical commitment to significance tests forces us to rephrase good questions in the negative, attempt to reject those nullities, and be left with nothing we can logically say about the questions." (Killeen, 2005, p.346)

Over the past few years, questionable research and publication practices in Psychology and Communication Science have been heavily debated (e.g., Rosenthal, 1979; Ferguson & Brannick, 2012; Simmons, Nelson, & Simonsohn, 2011; also see *Introduction* to this special issue). In the current paper, we go beyond the issue of favoring the publication of studies that find evidence for significant differences – publication bias – to discuss statistical approaches that may reduce the potential for such outcomes. Publication bias is the result of editors and journal policies to accept papers for publication that provide 'support' for the alternative hypothesis (H_a) rather than those that provide 'support' for the null hypothesis (H_0), due to the rationale behind traditional frequentist testing. This process can create a state of affairs in which published studies do not represent the actual population of results in a scientific field and may create a distorted and spurious perception of the strength behavioral theories actually have (Ferguson & Heene, 2012). Evidence for the presence of publication bias is, by now, considerable and fairly widespread (e.g., Ferguson & Brannick, 2012; Harrison et al., in press; Kepes & McDaniel, 2013). Theory-supportive results are far more prevalent in psychology and psychiatry than in the hard sciences (91.5% versus 70.2% in the space sciences, for instance, Fanelli, 2010), yet such a high success rate is impossible to achieve given the fact that many experiments generally have low power.

Publication bias occurs for several reasons. One may be because an apparently "true" theory may seem intuitively more valuable than one that appears falsified and is considered more interesting by journal editors. Second, researchers themselves may become (emotionally)

attached to their theories and engage in "obligatory replications" to further support their theories (Ioannidis, 2012). Third, and most critical for our argument in the current paper, weaknesses within null hypothesis significance testing do not properly allow for the falsification of theories. Null results are often seen as difficult to interpret and the product of low power, methodological problems in the study design or measurements issues, among others, and thus rejected. Such a rationale may drive some authors to add to a sample until their results become "statistically significant", no matter how small in effect size, creating a correlational pattern between sample size and effect size in published results (Fritz, Scherndl, & Kühberger, 2012). Consequently, such studies are difficult to replicate, given that the initial results were the product of low power, not robust results (Schooler, 2011). Thus, traditional null hypothesis significance testing (NHST, also called frequentist statistics[1]) may contribute to publication bias by being ill-suited to 'support' a null hypothesis when, in fact, a null result might be the best conclusion for a test of a particular theory.

Following the above, we argue that part of the issue in the occurrence of publication bias resides in traditional NHST, which forces scholars to binary decisions regarding the significance or non-significance of results. Whether or not a hypothesis finds support in the empirical data is based on the common acceptance of setting a hard boundary to force decisions to either accept or reject a hypothesis based on an arbitrary significance level in terms of a specific *p*-value <.05 (Bakker, van Dijk, & Wicherts, 2012; Wicherts, Borsboom, Kats, & Molenaar, 2006).[2] Such rigid yes-no decisions may have tempted some scholars to push their results into desired significance levels (Brown & Bobkowski, 2011). However,

[1] We refer in particular to the commonly used Fisher's *p*-value (or Null Hypothesis Significance Testing, NHST). In fact, Fisher's NHST does not have an alternative hypothesis (Ha) only the null hypothesis (H_0). It only tests the strength of evidence by calculating the probability of the observed value (or more extreme) than the observed value, based on the assumption that H_0 is true. The method in fact does not test H_0 against Ha.

[2] The pre-set α significance levels to which the obtained *p*-value is tested may also set to be more strict (e.g., $p < .01$), varying among disciplines, research designs, sample sizes, or measurement levels. Then, similar reasoning regarding publication bias still holds.

measures can be taken to prevent publication bias through an alternative approach of statistical testing of hypotheses.

We propose the use of Bayesian Factors (BFs) to arrive at a more transparent and flexible way of presenting statistical support for the hypothesis under consideration in comparison to others. With BFs one has to specify expectations before analyzing the data because the expectations are part of the input for the Bayesian analysis. This way, one has to be open and transparent about the specific hypotheses tested beforehand. The results of Bayesian testing then provide various degrees of support for any hypothesis, through Bayesian model selection of informative hypotheses (Hoijtink, 2011; Klugkist, Laudy, & Hoijtink, 2005) using Bayes Factors (Kass & Raftery, 1995). We illustrate this method in the current paper by re-analyzing three datasets of previously published data regarding 1) media violence effects with mixed interaction hypotheses; 2) mediated communication; and 3) visuospatial abilities across genders; also including explicit testing of null hypotheses.

To avoid any misunderstandings: in the current paper, we are only interested in re-analyzing the data as provided by the original authors; we will not evaluate the method or paper as such. The results are presented and discussed in view of possible interpretations. We highlight how a Bayesian approach is more open to a content-related argumentation of the level of support for a particular theoretical point of view, also if this is a null hypothesis. However, a Bayesian approach is not without fallacies. That is, the results of Bayesian statistics can lead to instrumental optimization of Bayes factors (i.e., "BF-hacking") similar to the phenomenon of *p*-hacking in traditional NHST, especially when dichotomous cut-off values for evidentiary power (and thus "publishability") are set. In line with the famous quote of Rosnow and Rosenthal (1989, p. 1277), we argue that "God would love a Bayes Factor of 3.01 nearly as much as a BF of 2.99" (see section "*BF-hacking*"). Furthermore, not any BF-value can be taken as conclusive support without careful consideration of its interpretation,

especially when BF values are small. To illustrate that 'simply' gathering more data does not in itself resolve low BF values, we present the results of a simulation study (see section "*Sensitivity Analyses*"). When low BFs occur, we argue, researchers should carefully interpret the results and conceptually replicate the findings. When results of a previous study are used as input for the new analyses (called Bayesian updating) results might gradually lead to higher BF values and hence a stronger support for a theory. Alternatively, updating might lead to the conclusion that support for a theory remains weak or perhaps even fails in the end.

Before presenting our illustrative cases, we will outline the basics of a Bayesian approach in the next section.

Testing a Null Hypothesis

As widely discussed elsewhere and briefly summarized in the above, current journal policy and academic practice are prone to yielding publication bias. As argued, closely related to publication bias is the difficulty in testing a null hypothesis through frequentist statistics (NHST). Important to note is that rejecting H_0 does *not* imply that the alternative hypothesis is supported. Also, failing to reject H_0 does not imply support in favor of the null hypothesis. Unfortunately, non-significance is often misinterpreted as support for the null hypothesis (Bakker & Wicherts, 2011), which is abundantly illustrated in the *Journal of Articles in Support of the Null Hypothesis*: "offering an outlet for experiments that do not reach the traditional significance levels ($p < .05$)." (http://www.jasnh.com/). However, NHST is designed to reject the H_0, not to support it. Hence, a decisive test to support the H_0 and allowing to conclude that two means are similar (i.e., the chance that the means appear different is minimal), is lacking. For decades, behavioral scientists argue that H_0-testing is irrelevant because they believe that any hypothesis can be stated in terms of an alternative hypothesis (e.g., Cohen, 1994; Krueger, 2001; Nickerson, 2000). That is, they state that

'always, something is going on', resulting in a primary focus on testing predictions regarding differences between groups or conditions and neglecting null hypotheses.

In contrast, we argue that theoretically relevant null-hypotheses can be found increasingly in today's mediated society and following future new technology applications. An illustrative case is "Sweetie" - the virtual Philippine girl who 'caught' pedophiles online (Hare, 2014; Jovanovic, 2013; Author(s)). The thousands of pedophiles engaged in conversations with her did not notice that she is not a real girl. An older example is the now famous Turing Test (Oppy & Dowe, 2011; Turing, 1950; Author(s)), while contemporary examples can be found in a continued blurring of mediated and real lives (e.g., Facebook romances). Likewise, virtually created services (e.g., e-health coaches, e-therapy, social robots) will increasingly be applied in healthcare and service professions due to limited resources and aging (Author(s)). Such innovations in communicative acts undoubtedly underscore the relevance of testing null hypotheses based on the assumption that mediated communication may create similar outcomes as non-mediated interaction (also see our examples below). However, testing such hypotheses through frequentist statistics will never provide clear support in favor of a null hypothesis; the best NHST can offer is a failure to reject H_0. Therefore, we coin an alternative way of testing hypotheses that better fits the dynamics of contemporary science in view of the criticisms raised and which offers an appropriate way to test theory based hypotheses through Bayesian Factors, even if the theory is the null hypothesis.

Bayesian Approach to Hypothesis Testing

The Bayesian paradigm[3] offers a very different view of hypothesis testing than the commonly applied frequentist testing against the null (e.g., Kaplan & Depaoli, 2013; V.E. Johnson, 2013; Van de Schoot et al., 2014). Instead of setting hard boundaries to force a

[3] A full introduction to Bayesian statistics is beyond the scope of the current paper and we refer to Van de Schoot and Depaoli (2014) as a highly accessible start: http://www.ehps.net/ehp/index.php/contents/article/view/ehp.v16.i2.p75/26).

decision to either reject or fail to reject H_0, Bayesian analyses provide various degrees of support for the hypotheses under consideration. Therefore, one can compare the support in the data for a set of pre-specified hypotheses, for example the null hypothesis versus the alternative hypothesis. A Bayes Factor (BF) can be computed for any combination of hypotheses using, for example, the software package *BIEMS* (Mulder, Hoijtink, & Klugkist, 2010; Mulder, Klugkist, Van de Schoot, Meeus, Selfhout, & Hoijtink, 2009), but see also the software *BayesFactor* (Morey & Rouder, 2012) or *JASP* (Love et al., 2015). The result of testing hypotheses in these software packages are BF-values that represent the amount of evidence favoring one hypothesis over another. See Kass and Raftery (1995) for a general introduction to Bayes Factors, see Klugkist et al., (2005) for the computation of BFs between specific theory-driven hypotheses, and see van de Schoot et al. (2011) for an easy to read introduction and a comparison between BFs and NHST-results. Finally, see Romeijn and van de Schoot (2008) for a more philosophical discussion on Bayesian hypothesis testing.

When one hypothesis is tested against an alternative hypothesis[4] and the results indicate that BF = 1, the result implies that both hypotheses are equally supported by the data. However, when BF = 10, for example, the support for one hypothesis is ten times larger than the support for the alternative hypothesis. If the BF < 1, then, the alternative hypothesis is supported by the data. Balancing the differences in the degree of support can thus be used to judge the relevance of the null and alternative hypothesis in direct comparison (Johnstone, 1990). Sellke et al. (2001) showed that the BF is preferable over a *p*-value when testing hypotheses because *p*-values tend to overestimate the evidence against the null hypothesis.

Typically, the procedure of evaluating hypotheses consists of three steps. In the first step, the researcher should specify the set of hypotheses of interest using equality and inequality constraints. For example, H_1 (mean1 > mean2 = mean3), H_2 (mean1> mean2 <

[4] The default setting of the BIEMS software is a .5/.5 prior probability and cannot be changed, otherwise the method to calculate the BF is not valid (see Klugkist, Laudy, & Hoijtink, 2005).

mean3), or the null hypothesis H_0 (mean1 = mean2 = mean3). In the second step, each hypothesis is tested against the so-called *unconstrained hypothesis* (H_{unc}); a hypothesis without any restrictions, so not including any comparisons or assumptions. The BFs for each comparison (i.e., BF(H_1, H_{unc}), BF(H_2, H_{unc}) and BF(H_0, H_{unc}) should be >1 indicating a certain level of support for each hypothesis compared to the "empty" unconstrained-hypothesis. In the third step, the BF can then be computed between the hypotheses of interest, for example BF(H_1, H_2), by simply dividing the BF of H_1 by the BF of H_2. The underlying details of Bayesian techniques will not be further addressed in this paper as they are clearly described in, among several others, Hoijtink (2011) and for software specific specifications see Mulder et al. (2009; 2010).

The focus in the current paper is to present and discuss the results and implications of our re-analyses of previously published data which used frequentist NHST, now applying Bayesian model selection to test the same and additional hypotheses. In discussing the results, we highlight that Bayesian testing offers more balanced decisions illustrating our arguments.

Illustrative Samples: Bayesian Analyses of Previously Published Data

Study 1: Media Violence Effects

We are very grateful to the authors Mario Gollwitzer and André Melzer to support our endeavor in providing the data underlying their paper titled *Macbeth and the Joystick: Evidence for Moral Cleansing after Playing a Violent Video Game* (Gollwitzer & Melzer, 2012). Their main hypothesis concerns testing the phenomenon that people wish to cleanse themselves physically when their moral self has been threatened (i.e., the "Macbeth effect"). They argue that inexperienced players of violent video games may experience such a moral threat when they do play such a game, especially when the game involves violence against

humans. Experienced players may apply other strategies to alleviate any moral concerns, they argue.

Hence, their first hypothesis (indicated by H_1) reads: Inexperienced players feel *more morally distressed* after playing a violent game against humans (*GTA*) than after playing a violent game against objects (*FlatOut*), and experienced players feel less morally distressed no matter what kind of game they played. Subsequently, their second hypothesis (H_2) reads: Inexperienced players prefer more hygiene products (i.e., "moral cleansing") after playing a violent game against humans than after playing a violent game against objects, whereas this should not be the case among experienced (i.e., frequent) players. In our view, the second part of both H_1 and H_2 could also be formulated as an interesting null hypothesis, which we further discuss below.

To test the hypotheses, seventy students played one of two violent video games (involving humans vs. objects) and were then asked to select 4 out of 10 gift products, half of which were hygiene products, and complete a questionnaire. To measure *Moral distress*, a 5-item-questionnaire was applied (e.g., "Did your actions during the game give you a bad conscience?"). To measure *Moral cleansing*, the number of hygiene products selected were counted. In addition, measures were taken to assess *Game experience* through the mean index of four items referring to video game experience. Due to a skewed distribution, participants were dichotomized into *in*experienced (n=36) and experienced (n=34), according to the Median (1.5) (Gollwitzer & Melzer, 2012).

Results NHST vs. Bayes Factors

Original results. The original ANOVA-results as presented in Gollwitzer and Melzer (2012) indicated that the null hypothesis should be rejected for both main effects (p<.01) as well as for the predicted interaction effect, $F(1,66)=4.52$, p=.04, η_p^2 =.06. Post-hoc analyses revealed that, as expected, inexperienced players felt more distressed after playing a violent

game against humans than after playing against objects, $t(34)=-3.21$, $p<.01$, $d=1.08$, whereas no such difference was found for experienced players, $t(32)=-0.65$, $p=.52$, $d=0.22$. However, as described in the above, NHST can only indicate that the null hypothesis should be rejected while no direct evidence for the alternative hypothesis (i.e., H_1) can be provided.

Testing H_2, results of the ANOVA as presented in Gollwitzer and Melzer (2012) with the number of hygiene products (i.e., moral cleansing) as dependent variable yielded no significant main effects ($p \geq .12$), but a significant interaction effect, $F(1,66)=5.99$, $p=.02$, $\eta_p^2 =.08$. As expected, inexperienced players selected more hygiene products after playing *GTA* than after playing *FlatOut*, $t(34)=-2.03$, $p=.05$, $d=0.68$, whereas no such difference was found for experienced players, $t(32)=1.49$, $p=.15$, $d=0.51$. These results indicate that the null hypothesis for the main effects and the interaction effect for the experienced players could not be rejected. Therefore, a direct test of the null hypothesis for the experienced players would be of interest.

Bayesian model selection using Bayes Factors. Testing the same hypothesis H_1 with a Bayesian approach using the software BIEMS, which proceeds in steps (see above), yielded partly similar results: The main effects model (Model 1) showed a Bayes Factor (BF) of 1.72, indicating that there is 1.72 times more support for the hypothesis that participants playing the human subjects violent game (*GTA*) report more moral distress than those playing against objects (*Flatout*) compared to the unconstrained model. Model 2, testing the main effect of the less experienced vs. the more experienced players also supported the previous results in that the less experienced players, experienced more moral distress with BF = 3.68. Given the magnitudes of BF, there is more support for the impact of experience with playing games on moral distress (Model 2) than for the impact of the type of game being played (Model 1), yet all BF values are low.

The interaction hypothesis (in fact the authors' H_1 in full) was tested in Model 3 and resulted in BF = 5.10 compared to the unconstrained model. Thus, the interaction hypothesis received 5.10/1.72 = 2.96 times more support compared to Model 1 and 5.10/3.68 = 1.38 times more support compared to Model 2. This provides some direct support for the original conclusion that *less* experienced players who played *GTA* (against human subjects) experienced more moral distress than less experienced players who played *Flatout* (against objects).

In addition, the second part of the interaction hypothesis H_1 may be rephrased as a null hypothesis assuming no differences for the experienced players. The Bayesian analysis (Model 4) provides support for this H_0: *Experienced* players report the same level of moral distress whether playing against humans or objects by BF = 2.53. Thus, this analysis adds to the original in being able to provide some direct support for an assumption that was not tested in the original paper with frequentist statistics.

Testing the same H_2 from the original paper with a Bayesian approach resulted in a somewhat different picture: The first main effects model (Model 1) showed a Bayes Factor of BF = 0.21 (i.e., no support for the main effect of game-type on moral cleansing) and the second main effects model (Model 2) showed little direct support for the hypothesis that *less* experienced players show more moral cleansing than more experienced players (Model 2): BF = 1.54. Hardly any stronger support is reflected in BF = 1.62 for the interaction hypothesis (Model 3), stating that less experienced players show more moral cleansing when playing *GTA* than when playing *Flatout*.

As with H_1, the second part of the interaction hypothesis H_2 can be rephrased as a null hypothesis assuming no differences for the experienced players. However, no support is expressed through BF = 0.87 (i.e., BF < 1, see introduction above) (Model 4) for experienced players choosing the same amount of hygiene products across games. Thus, the Bayesian

result reaches a different conclusion for H_2 than in the original paper. While the original authors speculated that experienced players may have picked less hygiene products after playing *GTA* than after playing *Flatout* to not wash away their joy (see p. 1360, left column), they did not test this hypothesis. We further discuss this below.

We also explored two extra hypotheses: 1) more experienced players playing *GTA* choose more hygiene products than more experienced players playing *Flatout* or 2) vice versa. For the first, we found a Bayes Factor of 0.19, indicating more support for the unconstrained hypothesis than the hypothesis indicating a difference. For the second additional hypothesis, we found a Bayes Factor of 1.79, indicating little direct support for the hypothesis that the more experienced players playing *Flatout* actually choose *more* hygiene products than more experienced players playing *GTA*.

Discussion Study 1

We argue that the Bayesian re-analysis provides more nuanced support for the hypotheses, both as originally stated by Gollwitzer and Melzer (2012) and the null hypotheses. Based on the original results, the authors could only reject (or fail to reject) the null hypothesis (as is the premise of NHST) without direct support for their original and theory-driven hypotheses. Through Bayesian testing via model selection, we can now provide some direct support in favor of the hypothesis that less experienced players experience more moral distress when playing *GTA* against humans than when playing *Flatout* against objects. However, regarding the moral cleansing, the Bayesian results showed only slightly stronger support for the interaction hypothesis than for the main effect of being inexperienced. Hence, the evidence for less experienced players selecting more hygiene products when the game involved violence against humans compared to violence against objects, is not very clear and seems more strongly the result of level of experience rather than the humanness of the opponents in the game. It should be noted that the BF values obtained are relatively low and

therefore not very convincing. The theoretical implications for the phenomenon that *in*experienced players wish to cleanse themselves physically when their moral self has been threatened, especially when the game involves violence against humans (Gollwitzer & Melzer, 2012), may therefore be reconsidered in stating that moral cleansing among inexperienced players occurs irrespective of the humanness of the opponent.

The second part of both interaction hypotheses (H_1, H_2), in our view, reflected null hypotheses for the experienced players that were not tested in the original article. Here, the additional value of a Bayesian approach is evident: in the original article, the frequentist statistics (NHST) could not directly test null hypotheses while the Bayesian analysis could. The re-analysis with Bayesian procedures (i.e., Bayes Factors) however, provided some direct support in favor of H_0, namely that experienced players are equally distressed in both game-conditions (H_1, second part). However, no convincing evidence was found, through Bayesian analyses, for our added hypothesis that experienced players show equal levels of moral cleansing in both game-conditions but rather some support for a contrasting hypothesis. Actually, Figure 2 in the original paper shows that experienced players chose somewhat more hygiene products after playing against *objects* than against humans (i.e., in the opposite direction).

The theoretical implications for Gollwitzer and Melzer's (2012) claim that experienced players may apply other strategies to alleviate any moral concerns, may therefore be reconsidered. First, while the experienced players reported less moral distress than the inexperienced ones, irrespective of game-type, the question is whether the absence/low level of moral distress can be considered 'moral threat' at all, and thus, were there any moral concerns to alleviate through other strategies (than moral cleansing) among the experienced players? Second, because the added assumption of 'similar levels of moral cleansing for both game-conditions among experienced players', is *not* supported through Bayesian analyses,

the picking of hygiene products for the experienced players after playing against objects (rather than humans) may have not much in common with moral cleansing and asks for alternative explanations.

In all, it seems that relying on *p*-values to test hypotheses through NHST may lead to overstating the support for one's hypotheses while the Bayesian approach to the same data provided less convincing support through relatively low BF values. Furthermore, our Bayesian approach presented some challenges and refinements to the original results, implicating some interesting theoretical (re)considerations. It should be noted, however, that the BF values obtained are relatively low and warrant replication studies.

Study 2: Null Hypothesis Testing in Mediated Communication

We are very grateful to the authors Astrid Rosenthal-von der Pütten and Nicole Krämer for generously providing the data underlying their paper titled *Investigations on empathy towards humans and robots using fMRI* (2014). The study attempted to answer the question whether humans show an emotional reaction towards a robot and whether this reaction differs from those towards a human. The theoretical introduction seems to provide arguments for both H_0 and H_1, illustrated for example in "Although we assume that there might be a common basis for emotional responses and empathy towards humans and robots, there is also evidence that the perception of robots and humans lead to different activation patterns in fMRI studies." (p. 203). However, sufficient empirical evidence suggests that people respond socially and emotionally in similar ways as they to in human-human interaction when interacting with artificial entities such as avatars, mediated communication, virtual characters, and robots, as reported by Rosenthal-von der Pütten et al. (2014) as well as in our own research (Author(s)), though results regarding emotional responses to robots are scarce. The authors then conclude that "also for the domain of emotional reactions, similarities between HRI [Human-Robot-Interaction] and HHI [Human-Human-Interaction]

can be expected." (p. 204). Hence, in our view, this study reflects a typical null hypothesis ('robots evoke similar empathy as humans when treated badly'). Likewise, similarity in empathy for humans and robots is what the authors stress in their evaluation and press releases of the results. However, a non-significant test according to traditional probability testing cannot be taken as support for the null hypothesis (Gallistel, 2009; Rouder, Speckman, Sun, Morey & Iverson, 2009). Therefore, to conclude 'similarity' for humans and robots, we need a direct test of H_0 which can be done with Bayesian analyses.

The authors, in contrast, conclude their introduction with alternative hypotheses, summarized as follows:

H_1: Violent treatment of either human or robot will cause more negative (a) and less positive (b) emotions than affectionate treatment (and vice versa: less negative (c) and more positive emotions (d) after affectionate treatment of either one compared to violent treatment).

H_2: Humans in video clips will elicit stronger emotions (a), will be attributed more feelings to (b), and elicit higher empathy than robots in video clips, whether treated affectionately or violently (c).

While the authors combined self-report with a more objective fMRI- measure, we will only re-analyze the self-report data. To test their hypotheses, videos of humans interacting with either another human (HHI), or a robot (HRI; an animal robot Plebo), and a neutral box (excluded form subsequent analyses) were created (images provided in Rosenthal-von der Pütten et al., 2014). These 'interaction dyads' refer to the first experimental factor. The second experimental factor was created by treating the partner in either a violent or affectionate way. The research design was a 3 ('interaction dyads') x 2 (violent or affectionate) within-subjects design with positive and negative affect (PANAS), attribution of emotionality, and empathy (both self-construed scales) as dependent variables. The hypotheses are tested with the same respondents as in the fMRI-scanner: Fourteen healthy

volunteers (nine female, five male), aged between 20-30 years (M = 23.50, SD = 2.93). After the fMRI session, participants saw the videos again on a computer screen and completed a questionnaire.

Results NHST vs. Bayes Factors

Original results. The original article of Rosenthal-van der Pütten et al. (2014) reports results of two-factorial repeated measures ANOVAs to test effects of "interaction dyad" (human vs robot) and "treatment behavior" (violent vs affectionate), with the positive and negative PANAS-subscale as separate dependent variables. The main effect of "treatment behavior" was significant for the Human-Robot Interaction (HRI) such that participants felt significantly more positive after the robot being treated affectionately by the human (in de video clip) than after the violent treatment in HRI, ($F(1,13) = 24.462$, $p < .001$; $\eta_p^2 = .653$). However, no such main effect was found for the Human-Human-Interaction (HHI), but rather a significant interaction effect of "treatment behavior" with "interaction dyad", ($F(1,13) = 8.822$, $p = .011$; $\eta_p^2 = .404$). While the authors conclude that these results indicate that participants felt most positive after the video showing friendly interactions with robots and least positive after the videos showing violent interactions with robots, the results might also be interpreted as follows: Violent treatment of either human or robot will cause less positive emotions than affectionate treatment of either. But, again, with NHST the only hypothesis tested is whether either H_0 is rejected or H_0 could not be rejected, hence limiting conclusions. A similar ANOVA was conducted with negative affect as dependent variable, resulting in participants reporting significantly more negative feelings after the violent treatment than after the friendly treatment of either robot or human, (HHI: $F(1,13) = 5.985$, $p = .029$; $\eta_p^2 = .315$; HRI: $F(1,13) = 73.688$, $p < .001$; $\eta_p^2 = .850$) (i.e., supporting the original authors' hypothesis H_{1a}). No main effect for "interaction dyad" (human vs robot) and no interaction effect occurred.

To test H_2, the authors conducted two separate one-way repeated measures ANOVA's with "interaction dyad" as independent variable and either one of the dependent variables empathy, attribution of feelings, and negative evaluation of the video clip. No main effects were found on empathy or the attribution of feelings, but a significant effect of "interaction dyad" on negative evaluation of the video clip was found. The human-human interaction was evaluated more negatively than the human-robot interaction, ($F(1,13) = 14.304$; $p = .002$; $\eta_p^2 = .524$; details in Rosenthal-van der Pütten, 2014). These results seem to indicate that human-human interactions elicit stronger emotions – in the form of a more negative evaluation of the video clip – from participants than human-robot interactions. But, again, with NHST conclusions are limited to either H_0 is rejected or H_0 could not be rejected.

Bayesian model selection using Bayes Factors. In our Bayesian re-analyses using Bayes Factors, we first tested the hypotheses as stated in the introduction (i.e., focused at directly testing the hypotheses as stated by the authors), then followed by our analyses to directly test the null hypotheses of interest.

Testing each of the sub-hypotheses separately through BIEMS, Bayes Factors showed that there is some direct support for H_1 (a, b, c, and d). Participants felt more negative emotions after viewing violent treatment of either human or robot, than after affectionate treatment (BF = 3.97). Participants also felt less positive emotions after viewing violent treatment of either human or robot, than after affectionate treatment (BF = 3.50). Bayes Factors showed less direct support for the second hypothesis. A Bayes Factor of 1.85 shows that there is some support for the hypothesis that human-human-interaction received *stronger* negative evaluations than human-robot-interaction when treated violently (however, note that a different variable has been used here; not the PANAS as in H_1). With regard to empathy, a Bayes Factor of 1.89 shows that there is some support for the hypothesis that participants felt

more empathy watching human-human interaction than robot-interaction. The model testing attribution of feelings has a BF of approximately 1.00.

As argued in the above, contemporary theorizing regarding affective responsiveness toward virtual others (i.e., mediated communication like robots) gives ground to expect that participants would report *similar* levels of positive and negative feelings toward both HHI (human) and HRI (robot), both when treated affectionately or violently. Therefore, we tested this H_0-assumption through Bayes.

The models testing the H_0-assumption for the violent interactions all have Bayes Factors ≥ 1.5, indicating there is a little more direct support for the null hypotheses than for the alternative ones. In other words, the assumption that participants report similar levels of positive and negative feelings in response to both humans (HHI) and robots (HRI) (in the affectionate condition), receives some support but not very convincingly. The BF values are rather low.

Furthermore, the models testing the H_0-assumption for the affectionate interactions all have Bayes Factors ≤ 0.3, indicating there is more support for the alternative hypotheses than the null hypothesis. In other words, the assumption that participants report similar levels of positive and negative feeling in response to both humans (HHI) and robots (HRI), when being treated affectionately, is *not* supported by the data.

To further explore what we found in the data, we tested whether participants experienced more positive or negative feelings in response to humans or robots. With regard to negative emotions, the model testing the hypothesis that participants experienced more negative emotions in response to humans being treated affectionately than in response to robots being treated affectionately was somewhat supported by the data (BF = 2.01). Looking at positive emotions, the model testing the hypothesis that participants experienced more

positive emotions in response to humans being treated affectionately than in response to robots being treated affectionately was also somewhat supported by the data (BF = 1.95). For hypothesis 2, we also tested the null hypothesis directly for each of the sub-hypotheses (H_{2a-c} separately). We found no support for all three sub-hypotheses. In each case, the BF was < 1.00. Thus, Bayes factors for each sub-hypothesis stating the null were *smaller* than for the original hypotheses predicting stronger reactions in response to the human than toward the robot.

Discussion Study 2

While the authors presented arguments for both hypotheses of differences between group means as well as null hypotheses, they opted for the first. However, they did not find support for rejecting the null hypotheses through the traditional frequentist statistics, which they used to test the hypotheses, and thus concluded that similar responses were found toward humans and robots. With our Bayesian approach, we could clearly add to these interpretations by providing results of testing both types of hypotheses against each other. It turned out that testing each of the sub-hypotheses separately through Bayes Factors, provided sometimes a little more support for the original hypotheses of differences in response toward humans and robots and in some cases some direct support for the null hypotheses of no difference. However, the results of BF-models directly testing the null effects on negative evaluation of violent interactions, empathy and attribution of feelings, indicated that no hypothesis was supported by the data. In all, although direct testing of the null hypotheses was possible through a Bayesian approach, the BF values are quite low and thus cannot count as convincing evidence.

An argument might be the rather small sample size which limits the power of probability testing. Important to note in this respect is that Bayes is less sensitive to sample

size and actually provides more accurate estimates with small sample sizes as well (Lee & Song, 2004; also see section "Sensitivity analysis" below).

When testing hypotheses related to users responding socially and emotionally to a robot (or computer) in quite the same way as they do to real humans, research needs to obtain support in the data for H_0. Likewise, virtual reality and augmented reality applications currently raise similar null hypotheses of interest. In other words, testing support in favor of H_0, rather than its rejection, is then the central theoretical quest which can be much better addressed with Bayesian statistics than with traditional frequentist statistical testing as illustrated in the above example.

Study 3: Visuospatial Abilities Across Genders

We are very grateful to Chris Ferguson for generously providing the data underlying his paper titled *Gender, video game playing habits and visual memory tasks* (2008). This paper provides an interesting case for our argument because the theoretical introduction seems to argue toward a H_0, but then results in a H_1. That is, the author debates results found thus far that men and women would innately differ in visuospatial abilities. In contrast, the author argues that differences in visual memory can be better explained by object familiarity (i.e., learning history). The paper further examines the role of prior training for visuospatial tasks through video gameplay. In concluding the introduction, the hypotheses assume domain-specific gender differences in visuospatial abilities (H_1) and (male) gaming experience as explaining these differences (H_2). Given the ambiguous status of the theoretical arguments, testing H_0 against alternative hypotheses is a challenge for a Bayesian approach because each can be addressed in direct competition.

To test the hypotheses, 72 college students completed the Rey complex figure test (Meyers & Meyers, 1995) to assess visual memory and perceptual organization (by drawing a complicated, abstract figure from memory). In addition, to measure visual recall, they were

asked to draw six objects, selected from pilot testing them as either male objects (revolver and video game controller), female (brassier and make-up compact), or neutral objects (bicycle and eyeglasses). Total drawing ability is assessed as a composite score of the Rey test and drawing score for the six exemplars. Furthermore, measurements for video game playing habits in general as well as for playing violent video games were based on Anderson and Dill (2000).

Results NHST vs. Bayes Factors

Original results. Before presenting the results of our re-analyses, we present the original results from the paper. Results of a MANOVA analysis indicated a significant main effect for gender on the Rey complex figure test, using Wilk's Lambda, $F(4, 67)=10.38$, $p < .001$; $d= .80$). However, subsequent univariate analyses showed that males and females did not differ on the Rey test, $F(1,70)= .00$, $p=.99$; $d<.001$) with similar means for males ($M=47.24$; $SD=13.32$) and females ($M=47.24$; $SD=11.49$). Hence, this asked for a direct test of H_0, which we can do through Bayes below.

Support for hypothesis H_1 stating that visual memory is domain specific, was gained through the univariate analyses showing greater visual memory recall among males for "masculine" items, $F(1,70)=14.63$, $p<.001$; $d=.93$; for "neutral" items, $F(1, 70)= 4.50$, $p = .037$; $d=.52$), while females demonstrated greater visual memory recall for "feminine" items, $F(1,70)= 4.41$, $p= .039$ $d=.49$).

The second hypothesis was tested including control for gender differences, through two stepwise regressions. The results of the first regression, with general video game playing habits and gender entered as predictors of visual memory (total score for drawing tasks), were statistically significant, $F(1,70)=8.51$, $p \leq .01$, indicating a positive predictive relationship, $R=.33$, $R^2=.11$. However, standardized coefficients (β or beta-weight) indicated that only

video game exposure (β=.33; partial *r*= .33) was a significant predictor of visual memory, not gender.

Results model selection using Bayes Factors. Testing each sub-hypothesis of the authors' hypothesis 1 through Bayes, we found similar results as the original study. However, Bayesian results added some *direct* support for the hypothesis stating *no* gender differences on the Rey task (BF = 3.09), and also added some direct support for males outperforming females for both the male and neutral exemplars (BFs are 2.01 and 1.94, respectively), while the females outperformed males for the female exemplars (BF = 1.95). Overall, however, the Bayes Factors are not very high and thus support for the theoretical assumptions is limited. To test each sub-hypothesis of the authors' hypothesis 2 through Bayes, we used several multiple regression analyses with the following inequality constraints: the effect of either *general* gaming experience, or experience with *violent* games > 0, and the effect of gender = 0. With regard to *general* gaming experience, the Bayes Factor was 11.42 indicating direct and clear support in favor of the hypothesis. With regard to *violent* gaming experience, the Bayes Factor was even 13.11, also indicating clear support in favor of the hypothesis (i.e., that gaming experience makes a difference but not gender). These two results show the added value of Bayes; we can now not only conclude that we were unable to reject the hypothesis that gender is *not* a predictor of visual memory after controlling for (violent) gaming experience, but we can also conclude that there is clear and direct support in favor of the null hypothesis that gender plays no role in visuospatial ability after controlling for (violent) gaming experience.

Discussion Study 3

The original paper argued that the differences between males and females in visuospatial abilities as found thus far could be the results of object familiarity through a specific learning history. While the authors formulated hypotheses of differences in gender-

specific domains, we reasoned that a null hypothesis could likewise have been formulated. With our Bayesian approach we could address each in direct competition. Where the authors did not find significant differences between males and females for the Rey test of visuospatial ability, testing H_0 through Bayes provided direct support that genders did not differ on the Rey test. In addition, in accordance with the original frequentist testing of H_1, Bayes Factors supported the domain-specific gender differences in visual memory recall. Yet, the BF-values for H_1 were not very strong.

In a second hypothesis, the role of 'prior training' through habitual video gameplay was tested and Bayes factors supported and complemented the original findings. The Bayesian analysis provided clear and direct support in favor of the null hypothesis that males and females do not differ in visuospatial ability when prior training through (violent) video gameplay (holding primarily for males) was controlled for. While the authors of the original paper could only test the alternative hypotheses, our Bayesian approach provided an important extra. In fact, the BF-values obtained for our H_2 reformulated as a null hypothesis were quite stronger than any of the aforementioned BF-values in our sample studies. While an absolute interpretation of BF-values or setting cut-off values is not desirable (see section "BF-hacking" below), low BFs should be interpreted with caution (see section "Sensitivity Analysis").

In sum, the re-analysis of this study through Bayesian statistics adds to the original results in further specifying and finding direct support for the lack of gender differences in visuospatial abilities after controlling for 'training' through (violent) video gameplay.

BF-Hacking

An important question remains unanswered, which is how to interpret the obtained BFs in the studies above? While they do provide nuanced insights in weighing the various test results against each other, when BF values are relatively low, it may be unclear to what

degree they can be used to 'accept' H_0. Stated differently, when is the BF high enough in favor of H_0 to conclude that H_0 is supported by the data? To assist researchers with interpreting BFs, several scholars have proposed to apply cut-off values. However, Bayes Factors are not immune to a phenomenon similar to *p*-hacking, in which researcher degrees of freedom can be used to promote Bayes Factors which are theory supportive, which we call BF-hacking (cf. Simonsohn, 2014).

Questionable research practices (QRPs), typically designed to convert undesirable null results to theory supportive statistically significant results, are now known to be common (e.g., Fang, Steen, & Casadevall, 2012; John, Loewenstein, & Prelec, 2012). To the extent that BFs remain dependent upon group mean differences and sample size dependent standard error calculations, QRPs that influence traditional NHST will also influence BFs. These would include convenient exclusion/inclusion of outliers, dropping DVs or unfavorable trials from analyses or reporting, methodological flexibility in the extraction of data, among others (see Wagenmakers et al., 2014). In this sense, Bayesian analyses can be thought of as an improvement for the *p*-hacking issue, but it is not immune to the application of QRPs.

To evaluate whether the BFs as reported in the studies above can be considered substantial, cut-off values for minimum BF-values are proposed by Jeffreys (1961), and Kass and Raftery (1995). They argue that BFs between 1-3 are considered "Not worth more than a bare mention" (Kass & Raftery, 1995, p. 777), while only values beyond 3 can count as some support, with values starting at 10 being "substantial" (Jeffreys) or beyond 20 or "positive" (Kass & Raftery). Only BFs >100 (Jeffreys) or 150 (Kass & Raftery) are considered "decisive". More recently, in the software JASP (Love et al., 2015), one asterisk is used when the BF >10, two asterisks are used when the BF >30 and three asterisks are used when the BF >100. This is probably implemented to help users of the software to interpret the results, but

these values are probably not meant by Love and colleagues as strict cut-off values, like $p <$ .05 is nowadays often (mis)used.

We argue strongly against using unjustified cut-off values as decision rules within the Bayesian framework because this might result in similar 'hacking-behavior' as with *p*-values. That is, a BF of 3.01 should not be considered 'substantial' in comparison to a BF of 2.99, which would then be "not worth more than a bare mention" (cf. Kass & Raftery). Put differently, using the famous quote of Rosnow and Rosenthal: "[…] surely, God loves the .06 nearly as much as the .05" (Rosnow & Rosenthal, 1989, p. 1277), warning us that the *p*-value should not be used as a dichotomous decision tool, but as a probability measure. We argue that the same holds for the BF (i.e., God would love a Bayes Factor of 3.01 nearly as much as a BF of 2.99). Therefore, we feel that arbitrarily applying cut-off values, as a clear underpinning is lacking in our view, might result in similar misuse of BFs as with *p*-values.

To provide more insight in how BF-values might vary given a particular dataset, we conducted some simulation analyses on the data provided.

Sensitivity Analysis

To get a sense of how to interpret the BF values as found in the re-analyzed studies reported in the above, we conducted some simulation studies that we called 'sensitivity analyses' to illustrate what changes would occur in the BF value when sample size would vary. That is, we conducted post-hoc analyses to scrutinize the sensitivity of BF if the obtained sample sizes would hypothetically increase/decrease (keeping all other factors the same).[5]

[5] Note that we do not want to claim that BF values have frequentist properties or that small BF values can be justified with such a simulation study. Furthermore, the sensitivity analyses are conducted given the acquired data, the statistical model and tested hypothesis. Furthermore, in addition to scrutinize the sensitivity of BF for variations in sample size, we also varied the mean difference between the groups and the variance. Due to space limitations, the latter results are not reported, yet, they do not alter the picture as described here. These results can be requested by sending an email to the corresponding author.

For the simulation study we used the software *BIEMS*, where one can generate data sets with exactly the same descriptive statistics as are given in the input. Hence, we generated new datasets that are look-a-likes for each of the three original datasets and provide insightful results to the relative value of obtained BFs. For each, new datasets were generated through increasing/decreasing the sample size while keeping the other parameters similar as in the empirical data. For each newly generated dataset, BFs were computed. In so doing, answering the question how the obtained BFs can be interpreted becomes possible in a way that still respects the original dataset (i.e., by keeping the other parameters constant). The results of our analyses provide insights into what the BF could have been if, for example, the sample size of the original data would have been larger/smaller.

In Figure 1, the results of the sensitivity analysis are shown for the moral distress study showing what BFs are obtained when the sample size is increased/decreased under the hypothesis that experienced gamers report the same amount of moral distress for both types of games. On the y-axis, the BFs are presented and on the x-axis the variations in sample size for the generated data sets (given the original variance and mean differences between the groups). The red dot indicates the BF as we found in the original dataset.

The results show that the BF gets somewhat higher with increasing sample size, leveling out around 4.0 once the sample size per group is > 60. This shows that support for the hypothesis that experiencing the same amount of moral distress in both conditions would become a bit stronger as sample size increases, but it is still not very impressive. This highlights that for the specific data as received by the authors and the specified hypotheses, the BF could never have reached the cut-off values as suggested by several authors discussed in the previous section. The data would still have low evidential value. This might be inherent to particular types of studies that include more 'noise' than clean lab data with highly

simplified stimuli. Nevertheless, our analyses suggest that replication studies are needed that may further improve the study design, materials and measurements.

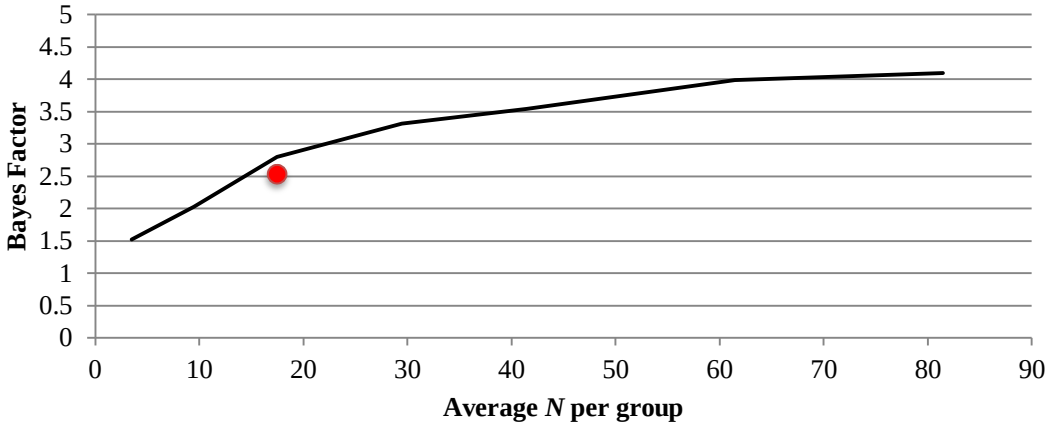

Figure 1. Bayes Factors for the generated data sets of the 'Moral Distress' study where the sample size has been increased/decreased, but with fixed mean difference and variance.

In Figure 2, the results are given for the newly generated datasets with varying sample sizes under the hypothesis that participants experience the same amount of either negative or positive feelings (i.e. left and right graphs, respectively) while watching humans or robots getting maltreated in the 'robot study'. As can be seen in Figure 2, the BF gets *lower when sample sizes increase*. The BF becomes even lower than 1.0 when the sample size is > 140 for negative feelings and when $n > 45$ for positive feelings. So, perhaps counter-intuitively, the collected data would show a decrease in support for the hypothesis of no differences among participants in both conditions if sample size would increase and group differences would remain equal.

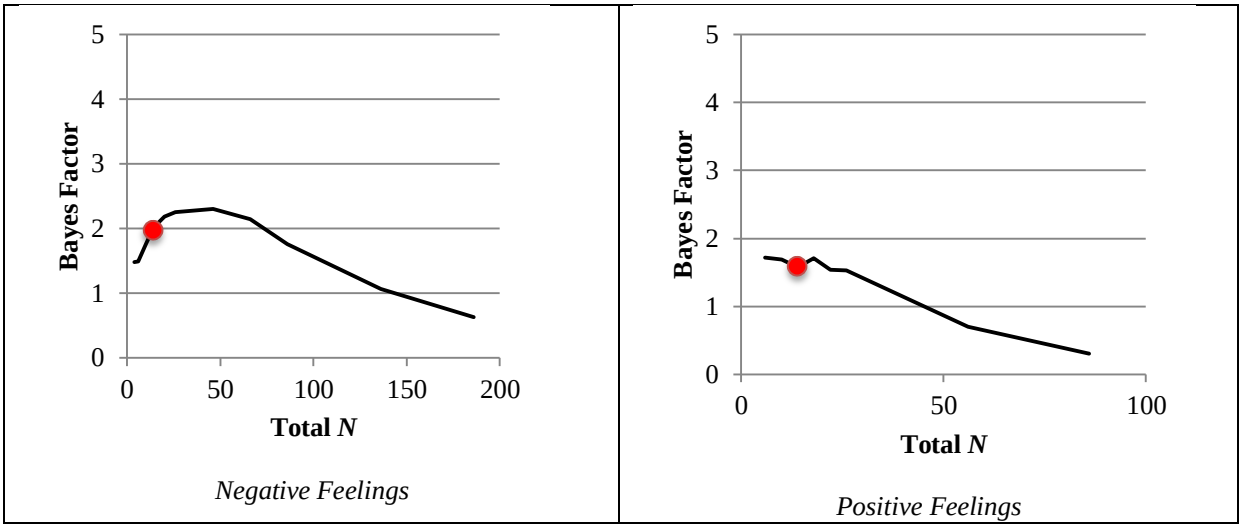

Figure 2. Bayes Factors for the generated data sets of the 'Robot Study' where the sample size has been increased/decreased, but with fixed mean difference and variance.

Figure 3 below shows the BFs for the newly generated datasets with varying sample sizes under the hypothesis that males and females have the same score on the Rey drawing task in the so called 'Visual memory study'. As sample size increases, so does the BF, reaching a BF > 5.0 once the sample is > 220. This shows that direct support for the hypothesis of no differences between groups, that is, men and women have the same score on visual memory, becomes stronger as sample size increases and group differences remain constant. However, given the current data, it would require an unrealistic sample size to obtain so called 'decisive' BFs > 10.

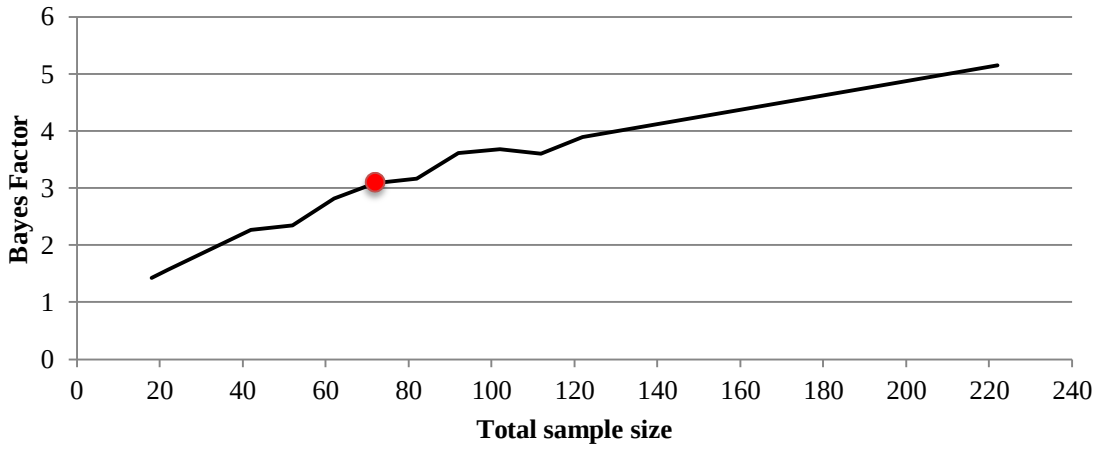

Figure 3. Bayes Factors for the generated data sets of the 'Visual Memory' study where the sample size has been increased/decreased, but with fixed mean difference and variance.

The simulation study demonstrated that the BFs as obtained in the original datasets are not likely to become much higher under different circumstances of the data. While Pericchi (2015) argues that Bayes factors (BF) may become more stringent as the sample size grows, the BF itself does not seem to become much higher as sample size grows. Note that this does not imply that the currently obtained low BFs should be attributed more weight than they have. However, it might mean that researchers in search of further evidence for the hypotheses proposed might not suffice in directly replicating the original experimental procedures and merely collecting a larger sample. This might lead to noisy samples where higher BF values are unlikely to be obtained. Instead, researchers might need to consider conducting conceptual replications under better controlled experimental conditions as to reduce noise.

In sum, the simulations suggest that the given datasets might have never reached the cut-off values as discussed in the previous section, even if the authors would have had much larger datasets. Please note, this is not to say that the obtained BFs are sufficient, yet, that given the stimuli and experimental procedures at hand, not much better results may be expected if larger samples are collected. This implies that currently there is only weak and sometimes no support for the theories under investigation and better controlled replications are needed to find out if evidence slowly accumulates over time in favor of the tested theory, or that in due time the theory needs to be discarded.

Even though the sensitivity analyses reflect somewhat critically on the strength of the data at hand, we argue that setting strict cut-off values for BFs is undesirable and would perhaps even result in publication bias (i.e., discarding studies with BFs below 10). While

small BF values should not be over-interpreted, they should at the same time not be ignored because they may play an equally important role in cumulative knowledge building as high BF values do.

General Discussion

The issue of publication bias in the context of null hypothesis significance testing (NHST) is, by now, well established across multiple research fields and multiple domains (Fanelli, 2010; Ferguson & Brannick, 2012; Nosek, Spies, & Motyl, 2012). NHST has already often been identified as problematic given the difficulty NHST has in offering support for null hypotheses (H_0), with concomitant difficulty in publishing such papers. This state of affairs has resulted in common questionable research practices (QRPs, see John, Loewenstein, & Prelec, 2012) through which scholars, often acting in good faith, purposefully seek out publishable "statistically significant" results. This developing culture of significance chasing ultimately does little to advance science, as the population of published studies may not reflect the true state of affairs in a given science. Were it possible to directly find support in favor of null hypotheses (H_0), this state of affairs might change. Under such a statistical system, null results could be worthwhile and potentially publishable. Indeed, "testing the null" would become an easier prospect. Likewise, in scientific fields related to media effects, statistical significance chasing under NHST may create researcher expectancies and cultural biases in which scholars may find what they are looking for. However, as we have illustrated in our examples in the current paper, it is just as viable to hypothesize that gameplay *does not* affect moral distress as much as it is to hypothesize that it does (i.e., for the inexperienced gamers). Also, the analyses in this paper indicated that interacting with robots may have similar outcomes as interacting with humans in some contexts but not in others, and gender may not make a difference for visuospatial abilities when taking one's learning history into account. Clearly, today's mediated environments and

new communication technology bring up many of such hypotheses (e.g., robots, social media, e-learning and e-therapy, augmented reality and virtual environments). If we are unable to test these ideas meaningfully, the entire scientific enterprise is difficult to interpret and may suffer from biases created by the 'need' to *proving* null hypotheses to be *not* 'true'.

With this in mind, we sought to examine a Bayesian approach to hypothesis testing as a new avenue for media studies. Under a Bayesian approach it is possible to suggest and test multiple hypotheses directly and in comparison to each other, including the null hypothesis, and offer varying degrees of support for each. Such an approach offers more nuanced and meaningful results in regards to competing hypotheses and it may be possible to get a clearer picture of the degree to which given data support one hypothesis over the other. Across a research field, this may also help to prevent bias in meta-analyses to the extent such analyses rely spuriously on published, positive findings and an unbiased sample of null studies may be difficult to locate (Ferguson & Brannick, 2012). However, as said, we should keep in mind that not being able to reject the null hypothesis in commonly used frequentist statistics (NHST) does not imply that the H_0 finds support in the data as is often erroneously inferred. To test a null hypothesis, or find support for a H_0, a direct comparison of H_0 versus meaningful alternative hypotheses through Bayesian statistics is necessary.

In the current paper, we reanalyzed three studies with datasets provided by the original authors. All three papers appeared to contain null hypotheses, some more explicit, others hidden within interaction analyses. With a Bayesian approach we were able to specify all individual sub-hypotheses and test them in a meaningful way. In some cases, our results found clearer support for the authors' original hypotheses than their use of NHST did, in other cases the evidence appeared weaker or not so clear. In most cases, the Bayesian reanalyses revealed rather low BF values which cannot be taken as convincing support and indicate that the originally obtained *p*-values overestimated the evidence against the null. We

have shown that a Bayesian approach (1) may lead to different conclusions than NHST, (2) provide appropriate ways to directly test (implicit) null-hypotheses, and (3) show that despite producing significant *p*-values, studies may have only small evidential value.

Often, researchers are very creative in translating their specific expectations to fit with traditional NHST. Therefore, it is sometimes difficult to find a straightforward relationship between the expectations as posed in the introduction section and the (null) hypotheses that should also be tested. The results then do not provide a direct answer to the research question at hand and in the discussion section, reseachers are again very creative to translate their result of NHST back to the original research question. Researchers then go back to the descriptive statistics, such as means, to answer the research question instead of correctly interpreting the NHST results or, alternatively, test the H_0 directly. This holds especially when the results are a bit counterintuitive or if many tests have been used (van de Schoot, Hoijtink, Mulder, Van Aken, Orobio de Castro, Meeus, & Romeijn, 2011). Our opening quote of Killeen (2005, p.346) could not have been more appropriate.

Through the Bayesian approach, such practices can be avoided as one has to specify each (sub)hypothesis in detail before looking at the data. Each hypothesis should reflect theory and can be based on inequality constraints reflecting statements like 'higher' or 'smaller', but can also consist of a typical null hypothesis assuming no differences between groups. For the Bayesian approach, it does not really matter which hypothesis is tested as long as it reflects theory. The toolbox of Bayesian statistics is therefore much more flexible than the frequentist toolbox (for NHST), where the logic always is to test a hypothesis by *rejecting* the hypothesis of 'nothing is going on' (H_0) and then conclude that 'something is going on' (but we don't know what). Of course, within the frequentist toolbox various suggestions have been coined to test the null hypothesis, also in the area of communication studies (e.g., Levine, Weber, Hullett, Park, & Massi Lindsey, 2008; Levine, Weber, Park, &

Hullett, 2008) or to change the alternative hypothesis into an inequality constrained hypothesis using contrast testing (e.g., Rosenthal, Rosnow & Rubin, 2000) or using bootstrapped methods (e.g., Silvapulle & Sen, 2004; van de Schoot, Hoijtink & Dekovic, 2010). However, these approaches are still limited by the focus on providing support for *rejecting* the H_0 rather than finding *direct support* for any hypothesis. Moreover, none of these methods are as flexible and easy to interpret as the Bayesian methods.

In the following, we will discuss some issues with a Bayesian approach that may be seen as problematic or limitations. Typically, Bayesian statistics do not have dichotomous benchmarks like BF > 3, or some even suggest BF > 10, to decide whether a study "worked" or not, and whether it is "publishable" or not. We believe setting a cut-off value would destroy the beauty of Bayesian testing and may lead to BF-hacking. The whole idea of providing various levels of direct support for various hypotheses that can then be compared directly in order to balance one's interpretation of the results, is an important feature of Bayesian testing that we should be aware of. Of course, low BF values might indicate only weak support for the hypothesis under consideration and replication of the results is still needed (Asendorpf et al., 2013) to rule out other explanations, weak data or unconsidered hypotheses. Furthermore, in some fields it might be harder to obtain large BF values than in others, especially when more natural materials or circumstances are examined as compared to 'clean' lab stimuli. To illustrate how the obtained BF-values would vary if certain model parameters would vary, we conducted so called sensitivity analyses to simulate what would happen if, for example, sample size would increase. It showed that the BF values could hardly get any larger in the given datasets. Therefore, the originally obtained *p*-values seemed to overestimate the evidence against the null and the results have low evidential value. Replication studies are needed to further test the theories at hand and to rule out other explanations. Furthermore, to improve the quality of the data, it might help to carefully

optimize the study design, materials and measurements. Other options for weak BF values are to combine several studies into a meta-analysis (e.g., Sutton and Abrams, 2001), or Bayesian updating where results of previous studies are updated with new data (e.g., van de Schoot et al., 2014). Science should be about accumulating evidence and one's individual results can hardly be approached as a final definite conclusion.

Moreover, the need of some to have a simple 'yes/no'-rule should be countered by more nuanced theorizing in which various options may have some support and other options are still open. An important aspect in the Bayesian approach is that one has to specify the hypothesis to be tested very precisely and therefore, no 'implicit' or 'hidden' assumptions can sneak through. Perhaps, more importantly, is the balanced output in terms of a certain weight of support for one hypothesis over another, yet, each hypothesis can be contrasted to several hypotheses through direct tests AND each receives a certain weight of support that can directly be compared to each other. Therefore, the dichotomous forced choice is avoided.
In line with our arguments in the previous section on possible BF-hacking, reviewers and journal editors should not take the magnitude of Bayes factors into account in their decision to accept papers. If they did, this could elicit publication bias in a similar way as the currently popular $p < .05$ benchmark. Then, scholars may resort to "BF-hacking", which does not seem to be more challenging than *p*-hacking (see Simonsohn, 2014, showing this). Simonsohn concludes that "Going Bayesian may offer some benefits, providing a solution to selective reporting is not one of them". We argue that by using Bayesian statistics, a researcher is required to be more open and transparent about the hypotheses tested and therefore the chances of BF-hacking are smaller than that of *p*-hacking. Yet, as a research community we should be cautious to not create new ways to introduce old publication bias.

Notes / Acknowledgments

The authors of the re-analyzed data reported in this paper are greatly acknowledged for generously providing their data for our analyses. Please, note that we do not evaluate the methods or quality of the research as such; we used these data sets for the current analyses only as interesting cases for our argument. The Netherland Institute for Advanced Studies (NIAS/KNAW) is acknowledged for granting the first author a fellowship allowing time to work on the current paper. In addition, we are very grateful to the reviewers of this paper, for their insightful and sharp feedback from which our paper benefitted a lot.

References

Anderson, C., & Dill, K. (2000). Video games and aggressive thoughts, feelings and behavior in the laboratory and in life. *Journal of Personality and Social Psychology, 78*, 772–790.

Sutton, Alex J. & Abrams, Keith R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Stat Methods Med Res., 10*, 277-303, doi:10.1177/096228020101000404

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H. and Wicherts, J. M. (2013), Recommendations for increasing replicability in psychology. *Eur. J. Pers., 27*, 108–119. doi: 10.1002/per.1919

Asparouhov, T. & Muthén, B. (2010). Bayesian analysis using Mplus: Technical implementation. Technical Report,version 3: http://statmodel.com/download/Bayes3.pdf

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666-678. doi: 10.3758/s13428-011-0089-5

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

Berger, James O. & Sellke, Thomas (1987). Testing a point null hypothesis: The irreconcilability of *p*-values and evidence**.** *Journal of the American Statistical Association***,** *82*(397), 112-122. DOI: 10.1080/01621459.1987.10478397

Brown, J. D., & Bobkowski, P. S. (2011). Older and newer media: Patterns of use and effects on adolescents' health and Well-Being. *Journal of Research on Adolescence, 21*(1), 95-113. doi: 10.1111/j.1532-7795.2010.00717.x

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Fanelli, D. (2010). 'Positive' results increase down the hierarchy of the sciences. *PLoS ONE, 5*, 4, e10068.

Fang, F. C., Steen, R. G., & Casadevall, A. (2012). Misconduct accounts for the majority of retracted scientific publications. Pro*ceedings of the National Academy of Sciences*, 201212247. doi:10.1073/pnas.1212247109

Ferguson, C.J., Cruz, A.M. & Rueda, S.M. (2008). Gender, video game playing habits and visual memory tasks. *Sex Roles, 58*, 279–286.

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120-128.

Ferguson, C. J. & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 550-556.

Fritz, A., Scherndl, T., & Kühberger, A. (2012, April). *Correlation between Effect Size and Sample Size in Psychological Research: Sources, Consequences, and Remedies*. 10th Conference of the Austrian Psychological Society, Graz, Austria.

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review, 116*(2), 439-453. doi:10.1037/a0015251; 10.1037/a0015251

Gollwitzer, M. & Melzer, A. (2012). Macbeth and the Joystick: Evidence for Moral Cleansing after Playing a Violent Video Game. *Journal of Experimental Social Psychology, 48,* 1356–1360.

Hare, E. (2014, May 23). D&AD Winners (Sweetie by Lemz and Terre des Hommes won in the White Pencil category of D&AD award competition for social adverts) [web log post]. Retrieved from: http://www.contagious.com/blogs/news-and-views/14215281-d-ad-winnners

Harrison, J. S., Banks, G., Pollack, J. M., O'Boyle, E. H., & Short, J. (2014). Publication bias in strategic management research. *Journal of Management*, doi:10.1177/0149206314535438

Hoijtink, H. (2011). Informative hypotheses: Theory and practice for behavioral and social scientists. Chapman and Hall/CRC

Ioannidis, J. (2012). Scientific inbreeding and same-team replication: Type D personality as an example. *Journal of Psychosomatic Research, 73*, 408-410.

Jeffreys, H. (1961). Theory of probability (3rd ed.). Oxford, UK: 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*(5), 524-532. doi:10.1177/0956797611430953

Johnson, V.E. (2013). Revised standards for statistical evidence, *PNAS, 110* (48), 19313-19317. Doi: 10.1073/pnas.1313476110

Johnstone, D. J. (1990). Interpreting statistical insignificance: A Bayesian perspective. *Psychological Reports, 66*(1), 115-121. doi: 10.2466/pr0.1990.66.1.115

Jovanovic, D. (2013, November 5). 'Sweetie' sting lures thousands of alleged pedophiles. *ABC News.* Retrieved from http://abcnews.go.com/WNT/video/meet-sweetie-virtual-girl-identified-1000-pedophiles-world-20797023

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*, 773-795. doi: 10.2466/pr0.1990.66.1.115

Kaplan, D., & Depaoli, S. (2013). Bayesian statistical methods. In T. D. Little (Ed.), *Oxford handbook of quantitative methods* (pp. 407–437). Oxford, UK: Oxford University Press.

Kepes, S., & McDaniel, M. A. (2013). How trustworthy is the scientific literature in industrial and organizational psychology? *Industrial and Organizational Psychology: Perspectives on Science and Practice*, *6*(3), 252-268. doi:10.1111/iops.12045

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16*, 345–353.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality constrained analysis of variance: A Bayesian approach. *Psychological Methods*, 10: 477-493.

Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Bayesian eggs and Bayesian omelettes. *Psychological Methods, 10*, 500-503. doi: 10.1037/1082-989X.10.4.500

Krueger, J. I. (2001). Null hypothesis significance testing. On the survival of a flawed method. *American Psychologist, 56*, 16-26. doi:10.1037//0003-066x.56.1.16

Lee, S.-Y., & Song, X.-Y. (2004). Evaluation of the Bayesian and Maximum Likelihood Approaches in Analyzing Structural Equation Models with Small Sample Sizes. *Multivariate Behavioral Research, 39*(4), 653 - 686.

Levine, T.R., Weber, R., Hullett, C.R., Park, H.S. & Massi Lindsey, L.L. (2008). A Critical Assessment of Null Hypothesis Significance Testing in Quantitative Communication Research. *Human Communication Research, 34*, 171–187. doi:10.1111/j.1468-2958.2008.00317.x

Levine, T.R., Weber, R., Park, H.S., & Hullett, C.R. (2008). A Communication Researchers' Guide to Null Hypothesis Significance Testing and Alternatives. *Human Communication Research, 34*, 188–209. doi:10.1111/j.1468-2958.2008.00318.x

Love, J., Selker, R., Marsman, M., Jamil, T., Verhagen, A. J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Rouder, J. N., Morey, R. D. & Wagenmakers, E.-J. (2015). JASP (Version 0.6.6) [Computer software].

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*(3), 679-690. doi:10.1037//0003-066x.56.1.16

Meyers, J., & Meyers, K. (1995). Rey complex figure test under four different administration procedures. *The Clinical Neuropsychologist, 9*, 63–67.

Morey, R. D. and Rouder, J. N. (2012) BayesFactor: An R package for computing Bayes factor for a variety of psychological research designs (available on the Comprehensive R Archive Network).

Mulder, J., Hoijtink, H., & de Leeuw C. (2012). BIEMS: A Fortran 90 program for calculating bayes factor for inequality and equality constrained models. *Journal of Statistical Software, 46*, 1-39.

Mulder, J., Hoijtink, H., & Klugkist, I. (2010). Equality and inequality constrained multivariate linear models: Objective model selection using constrained posterior priors. *Journal of Statistical Planning and Inference, 4*, 887-906.

Mulder, J., Klugkist, I., Van de Schoot R., Meeus, W., Selfhout, M., & Hoijtink, H. (2009). Bayesian model selection of informative hypotheses for repeated measurements. *Journal of Mathematical Psychology, 53*, 530-546. doi:10.1037//0003-066x.56.1.16

Muthén, L.K. and Muthén, B.O. (1998-2012). *Mplus User's Guide* (Seventh Edition). Los Angeles, CA: Muthén & Muthén.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301. doi:10.1037//1082-989X.5.2.241

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*(6), 615-631. doi:10.1177/1745691612459058

Oppy, G., & Dowe, D. (2011). The Turing test. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: The Metaphysics Research Lab, Stanford University, ISSN 1095-5054.

Pericchi, Luis (March 2015). The Chronicle of a Death Foretold Rejection Rule. Invited Contribution, *ISBA Bulletin*, *22*(1), p.7.

Romeijn, J. W. & Van de Schoot, R. (2008). A philosopher's view on Bayesian evaluation of informative hypotheses. In H. Hoijtink, I. Klugkist, & P. Boelen (Eds.). Bayesian evaluation of informative hypotheses (pp. 329-358). New-York: Springer.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge, England: Cambridge University Press.

Rosnow, R.L., Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276 – 1284.

Rosenthal-von der Pütten, Astrid M., Schulte, Frank P., Eimler, Sabrina C. , Sobieraj, Sabrina, Hoffmann, Laura, Maderwald, Stefan, Brand, Matthias, Krämer, Nicole C. (2014). Investigations on empathy towards humans and robots using fMRI. *Computers in Human Behavior, 33*, 201–212.

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*(2), 225-237. doi:10.3758/PBR.16.2.225; 10.3758/PBR.16.2.225

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437.

Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of *p* values for testing precise null hypotheses. *The American Statistician, 55,* 62-71. 10.1111/cdev.12169

Silvapulle, M. J., & Sen, P. K. (2004). Constrained statistical inference: Order, inequality, and shape constraints. London: Wiley.

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.

Simonsohn, Uri (2014, January 2), Posterior-Hacking: Selective Reporting Invalidates Bayesian Results Also. University of Pennsylvania, The Wharton School. Retrieved d.d. 21 September 2014 from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2374040

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433-460. doi:10.1093/mind/LIX.236.433

van de Schoot, R., and Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist, 16*(2), 75–84.

Van de Schoot, R., Hoijtink, H. & Deković, M. (2010). Testing inequality constrained hypotheses in SEM Models. Structural Equation Modeling, 17, 443–463.

Van de Schoot, R., Hoijtink, H., Mulder, J., Van Aken, M. A. G., Orobio de Castro, B., Meeus, W. & Romeijn, J.-W. (2011). Evaluating expectations about negative emotional states of aggressive boys using Bayesian model selection. *Developmental Psychology, 47,* 203-212.

Van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J. & van Aken, M. A. G. (2014). A gentle introduction to Bayesian Analysis: Applications to research in child development. *Child Development, 85*(3), 842–860.

Wagenmakers, E. (2007). A practical solution to the pervasive problem of *p* values. *Psychonomic Bulletin & Review, 14*, 779. doi: 10.1177/009365092019001003

Wagenmakers, E.-J., Verhagen, A. J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (in press). The need for Bayesian hypothesis testing in psychological science. In Lilienfeld, S. O., & Waldman, I. (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions.* John Wiley and Sons.

Wetzels, R., Raaijmakers, J. G., Jakab, E., & Wagenmakers, E. J. (2009). How to quantify support for and against the null hypothesis: A flexible WinBUGS implementation of a default Bayesian *t* test. *Psychonomic Bulletin & Review, 16*(4), 752-760. doi:10.3758/PBR.16.4.752

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