Journal of Articles in Support of the Null Hypothesis
Vol. 15, No. 2
 1539-8714
www.jasnh.com

Black sheep are not black in Wikipedia: Comparing descriptions of perpetrators in the language version of the perpetrator in-group to other (out-group) language versions

Aileen Oeberst
FernUniversität Hagen
Leibniz-Institut für Wissensmedien
Christina Matschke
Leibniz-Institut für Wissensmedien

People often evaluate in-group members, who deviated from socially accepted norms, more harshly than outgroup members who performed the same misconduct. The present paper investigates whether or not this so-called Black Sheep Effect also occurs in Wikipedia – a context that operates on strong non-evaluative norms and is the result of collaboration among diverse authors. To this end we conducted automatic text analyses for articles about $N = 149$ perpetrators (e.g., homicidal maniacs, terrorists, serial killers) and compared the relative use of negative emotion words (and anger, in particular) in in-group and out-group articles. With a Bayesian approach we found no support for the Black Sheep Effect, but much more support for the null-hypothesis.

Keywords: Black Sheep Effect, Wikipedia, collective memory, in-group bias, Neutral Point of View, norms

Corresponding author:
Prof. Dr. Aileen Oeberst, Institute of Psychology, University of Hagen, Feithstr. 188, 58084 Hagen, Germany,
Email: aileen.oeberst@fernuni-hagen.de

In March 2016 the United Nations published information about UN's blue helmets, who have been accused of sexual abuse on their missions.[1] This information lead to a wave of intense indignation in the nations involved in the blue helmet's mission: shocking negative behavior that is – disgusting as it is – frequently displayed in wars by all kind of men, was displayed by *one of us*. UN-secretary Atul Khare declared: "The horrible deeds of a few undermine the reputation of many."[2] In other words, the event was not only seen as negative behavior of individuals, but to shed a negative light on the whole group these individuals represent. Situations, where in-group members who act negatively are evaluated particularly harshly – even more so than members of another group (out-group) who performed the same misconduct, is referred to as the *black sheep effect* (BSE; Marques & Yzerbyt, 1988).

Why would we evaluate our own in-group fellows (compared to out-group members) especially harshly, when they act in a negative way? According to the social identity approach (Tajfel & Turner, 1979; Turner, Hogg, Oakes, Reicher, & Wetherell, 1987), people are generally motivated to view the groups they belong to (in-groups) positively because they contribute to their self-concept. In-group members that act negatively threaten the positive image of that group. Group members often deal with such threats by clearly distancing themselves from those, whose misconduct threatens the group and to frame them as *black sheeps*, who pose an exception to the otherwise proper group members. Thus, the black sheep effect serves the preservation or restoration of a positive representation of the in-group (Marques, Yzerbyt, & Leyens, 1988; see also Otten & Gordijn, 2014). In fact, the rejection of a negative in-group member has been shown to lead to a less negative evaluation of the group itself afterwards (van Leeuwen, van den Bosch, Castano, & Hopman, 2010).

According to the *Coping with Ingroup Deviance* model (CID; Otten & Gordijn, 2014), a BSE results if fully integrated group members (e.g., Pinto, Marques, Levine, & Abrams, 2010) intentionally (Braun, 2010; Braun, Otten, & Gordijn, 2009) deviate from a general humanity norm or a specific group norm (Marques, Abrams, & Serôdio, 2001) and are evaluated by people, who strongly identify with the group (Braun et al., 2009; Marques et al., 1988). The harsh evaluation of the deviant in-group members results from the fact that they do not act according to the expectations based on the in-group norm, which produces anger (Braun, 2010; Braun et al., 2009; van Prooijen, 2006) in the evaluator. Anger has been previously shown to mediate the relation between the deviance and the harsh negative evaluation by in-group members that characterizes the BSE (Braun, 2010; Braun et al., 2009; van Prooijen, 2006).

The BSE has been robustly and repeatedly documented (see Otten & Gordijn, 2014, for an overview). There are, however, contexts that might make the BSE less likely. The present research investigates Wikipedia as such a context. Two major reasons argue against a BSE in Wikipedia articles. First, Wikipedia operates on norms that aim at preventing bias. Specifically, Wikipedia urges authors to insert only information that is verifiable and from reliable sources[3]. Moreover, Wikipedia requires a neutral point of view and a factual presentation[4]. Wikipedia is a repository for *recognized* knowledge and not the

1 http://www.un.org/apps/news/story.asp?NewsID=53120 [May 16, 2016]; http://www.washingtonpost.com/sf/world/2016/02/27/peacekeepers/ [May 16, 2016]

2 https://www.tagesschau.de/ausland/blauhelme-missbrauch-101.html [May 16, 2016]

3 http://en.wikipedia.org/wiki/Wikipedia:Verifiability [May 16, 2016]

4 http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view [May 16, 2016]

place for own considerations[5]. Prior research has shown that these norms effectively guide authors' contributions (Forte & Bruckmann, 2008; Oeberst, Halatchlyiski, Kimmerle, & Cress, 2014; Viégas, Wattenberg, Kriss, & van Ham, 2007) and attest a high quality to Wikipedia articles (e.g., Giles, 2005). Also, research on other biases suggests that norms promoting an unbiased information processing may, in fact, reduce biases (e.g., Postmes, Spears, & Cihangir, 2001). Most notable here is an investigation on hindsight bias, which has been very robustly found in individuals' perceptions and judgments (Guilbault, Bryant, Brockingrouway, & Posavac, 2004; Roese & Vohs, 2012) but has been documented only for a small minority of Wikipedia articles (Oeberst et al., 2018). The second major reason for why we might not expect a Black Sheep effect in Wikipedia is that Wikipedia contains collaboratively written articles, that are socially negotiated among many authors. This allows a greater heterogeneity of viewpoints in the evaluators that describe potential black sheeps. Heterogeneity among collaborators in Wikipedia is especially likely, because it exists in language versions, not in country-specific versions (www.wikipedia.org). Thus, articles about perpetrators of one nation (e.g., U.S.-American) are not only written by in-group members, but also by out-group members (e.g., from the UK, but also Germany, Stvilia, Al-Faraj, & Yi, 2009). Such a heterogeneity may reduce or even eliminate biases (Schulz-Hardt, Frey, Lüthgens, & Moscovici, 2000; Vinokur & Burnstein, 1978).

Taken together, there is reason to suppose that Wikipedia is a context that might reduce or even eliminate the BSE. On the other hand, the BSE effect is a robust effect, and there is a controversy over the question whether people are able to fully control their biases (e.g., Fehr, Sassenberg, & Jonas, 2012; Macrae & Bodenhausen, 2000; Monteith, Sherman, & Devine, 1998). Moreover, some research hints towards potential group-based biases that occur in Wikipedia despite the norms of neutrality and the heterogeneity of authors. It has been shown that people prefer topics of the in-group (Hecht & Gergle, 2009) and provide more detailed elaborations on people from the in-group than the out-group (Callahan & Herring, 2011). In other words, it is likely that articles on perpetrators are, in the end, mostly written by in-group members, which would reduce the heterogeneity of the authorship after all and might therefore not counteract biases so much.

In sum, the literature about the BSE in Wikipedia provides support for both, the BSE-hypothesis, that is, a harsher evaluation of negative behavior when performed by an ingroup member than when performed by an outgroup member, as well as the null-hypothesis (i.e., no differences in the evaluation of negative behavior as a function of membership of the target person).

Method

We extracted articles about people, who intentionally, verifiably and severely deviated from general humanity norms (Otten & Gordijn, 2014). Specifically, we searched for perpetrators for whom articles existed in the language of the perpetrator's origin (i.e., where presumably more in-group members describe the perpetrator) and in another language (i.e., where presumably more out-group describes the perpetrator). We then compared the presentation of the perpetrator in the articles from the different language versions. Applying automatic text analyses we determined the percentage of negative emotion words in each article version. A BSE would be evident if the same person was

5 http://en.wikipedia.org/wiki/Wikipedia:No_original_research [May 16, 2016]

presented more negatively in the in-group article than in the out-group articles.

Material

We conducted an extensive search for people from the categories: homicidal maniacs, serial killers, sexual offenders[6], fraudsters, terrorists, and people from organized crime, and included all people for whom the following criteria were fulfilled. First, the in-group of the person had to speak one of the languages for which dictionaries with emotion words exist for the automatic text analysis tool we used (LIWC, see Analysis; included languages: Dutch, English, French, German, Italian, Portuguese, Russian, Spanish, Serbian, Turkish). Second, the negative act had to be officially verified (i.e., accusations alone did not count; see van Prooijen, 2006). Third, Wikipedia articles about the person had to exist – or Wikipedia articles about the event (e.g., the gun rampage), which, however included a paragraph about the perpetrator – in the language version of the in-group and in at least one out-group language version. This resulted in a total of $N = 149$ people ($n = 28$ homicidal maniacs, $n = 25$ serial killers, $n = 17$ sexual offenders, $n = 31$ fraudsters, $n = 26$ terrorists, $n = 22$ from organized crime). The list of people as well as the data, analyses and results can be retrieved from https://osf.io/2ukwn/.

In a next step, we determined which parts of the article were to be analyzed. We limited our analysis to the presentation of the perpetrator. Therefore, we excluded paragraphs about irrelevant aspects (e.g., inspired movies) but also about the specific deeds of that person because (a) the black sheep effect concerns the evaluation of the person and (b) the specific deeds likely vary in their emotional content (e.g., gun rampages being described with more negative terms than fraud). The classification of content was partly realized by two raters who examined the table of contents of randomly chosen articles (19 perpetrators, 184 classifications) and determined for each section whether it contained a presentation of the perpetrator or not. Their agreement was good (Cohen's $\kappa = .793$, $p < .001$). Disagreement was resolved by discussion and for the rest of the cases content classification was made by one rater only.

Resulting from this content classification task we had at least two texts per person – one from the in-group language version of Wikipedia and at least one from an out-group language version. However, in many cases we were able to retrieve articles from several out-group language versions, which were later averaged. All texts resulting from this step were subjected to automatic text analysis.

Analysis

As a criterion for the negative evaluation of a perpetrator within Wikipedia articles, we chose the percentage of negative emotion words in the presentation of the perpetrator. For the analysis, we made use of the automatic text analysis tool *Linguistic Inquiry and Word Count* (LIWC, Tausczik & Pennebaker, 2010). The tool is based on expert-defined and standardized dictionaries, in which words have been classified into different categories (e.g., negative emotion words such as "hurt", "nasty"). When analyzing text, the tool assesses the relative frequency of words that fall into each of the pre-defined categories (e.g., the

6 In the sexual offenders group we did not include people who also killed their victims. In the serial killer group, however, there were also people who did not only kill but also sexually abused their victims.

percentage of negative emotion words in the analyzed text). Therefore, the tool offers a quantitative and objective measure and it has been shown to be reliable and valid (see Tausczik & Pennebaker, 2010 for an overview). Moreover, it has been successfully employed in various research contexts such as motives (e.g., Schultheiss, 2013), personality (e.g., Fast & Funder, 2008), disorders (e.g., Wolf, Sedway, Bulik, & Kordy, 2007), but also in the context of emotional contents, which is most important for the present purpose (see Greving, Oeberst, Kimmerle, & Cress, 2018, for research into emotional content in Wikipedia articles about negative events). Moreover, due to the fact that dictionaries from different languages are available it enables inter-language comparisons.

We selected the following categories from the LIWC dictionaries for our analysis. First, we determined the percentage of *negative emotion words*, which was used for the presentation of the perpetrator. Second, as the black sheep effect is driven by anger as emotional process (Braun, 2010; Braun et al., 2009; von Prooijen, 2006), we also separately assessed the percentage of *anger*-related words, which is a sub-category of the negative emotion words. Third, we identified the percentage of *positive emotion words,* in order to test whether it is the general emotionality that accounts for the effect, or a decrease in positive evaluations, or specifically the harsh evaluation of perpetrators. A BSE would be displayed if a significantly higher percentage of negative emotion words and anger-related words were used to present perpetrators in the in-group version of Wikipedia articles compared to the out-group versions of Wikipedia articles.

Since there were arguments both, in favor of the BSE-hypothesis as well as against it (and thus for the null-hypothesis) we took a Bayesian approach as it provides insight into the strength of evidence for either hypothesis. All analyses were conducted with JASP (version 0.8.4).

Results

Main Analyses

Negative emotion words. We first averaged the percentages of all available out-group language versions in order to compare them with the in-group language version. We then conducted a Bayesian paired *t*-test testing the BSE-hypothesis ($M_{ingroupversion} > M_{outgroupversion}$, see Figure 1). As displayed in Figure 2, the analysis yielded strong support *against* this hypothesis, $BF_{+0} = 0.048$ (see also Jarosz & Wiley, 2014, for conventions regarding the interpretation of Bayes Factors) as the null-hypothesis was 20.98 times more likely given the data.

Anger. We ran the same two Bayesian paired *t*-tests for the percentage of anger-related words. The test of the BSE-hypothesis ($M_{ingroupversion} > M_{outgroupversion}$, see Figure 3), again, was not at all supported by the data, $BF_{+0} = 0.033$, indicating that an alternative to this hypothesis was 30.64 more likely given the data (see Figure 4). Thus, we found, again strong to very strong evidence against the BSE-hypothesis, which can also be derived from the descriptive pattern from Figure 3.

Exploratory Analyses

Since we did not specify any hypotheses for the following analyses, we only tested for (non-directional) differences (i.e., $M_{ingroupversion} \neq M_{outgroupversion}$).

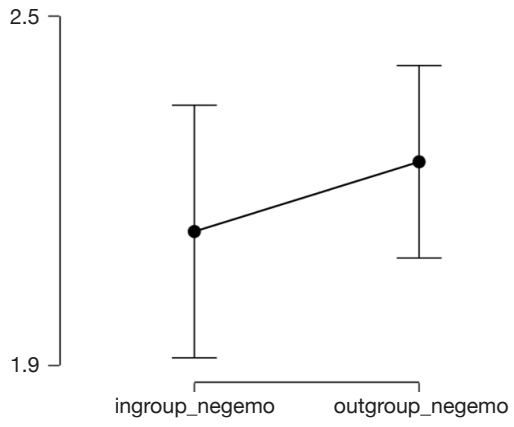

Figure 1. *Percentage of negative emotion words as a function of language version of the Wikipedia article*

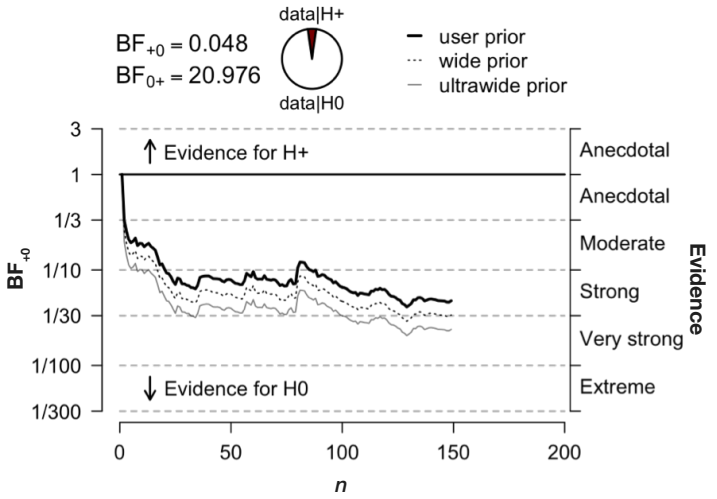

Figure 2. *Sequential evidence against the BSE-Hypothesis regarding percentage of negative emotion words ($M_{ingroupversion}$ > $M_{outgroupversion}$) as a function of prior (Bayesian paired t-test)*

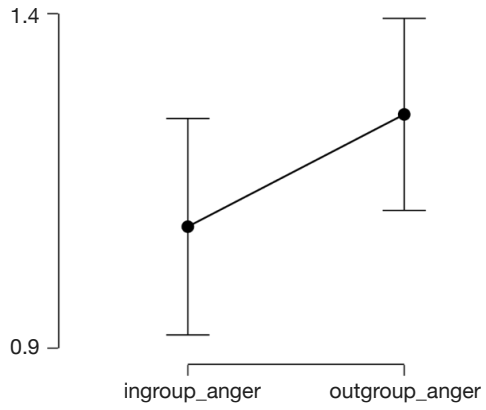

Figure 3. *Percentage of anger-related words as a function of language version of the Wikipedia article*

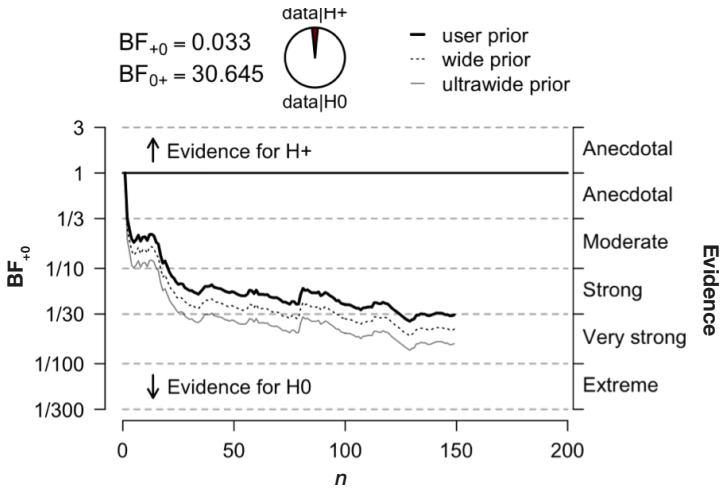

Figure 4. *Sequential evidence against the BSE-Hypothesis regarding percentage of anger-related words ($M_{ingroupversion}$ > $M_{outgroupversion}$) as a function of prior (Bayesian paired t-test)*

Positive emotion words. A Bayesian paired *t*-test comparing the percentage of positive emotion words yielded, again, more support for the comparability between language versions rather than for differences with regard to emotionality, BF_{01} = 5.747, $M_{ingroupversion}$ = 1.431, *SD* = 1.204, $M_{outgroupversion}$ = 1.315, *SD* = .0584.

Self-focus. We tested whether previous findings of more elaborated accounts of in-group topics compared to out-group topics (Hecht & Gergle, 2009; 2010) could be replicated with our data. A Bayesian *t*-test provided decisive evidence for this pattern, $M_{ingroupversion}$ = 1774.0, *SD* = 2926.2, $M_{outgroupversion}$ = 674.5, *SD* = 674.5, BF_{10} = 14007.

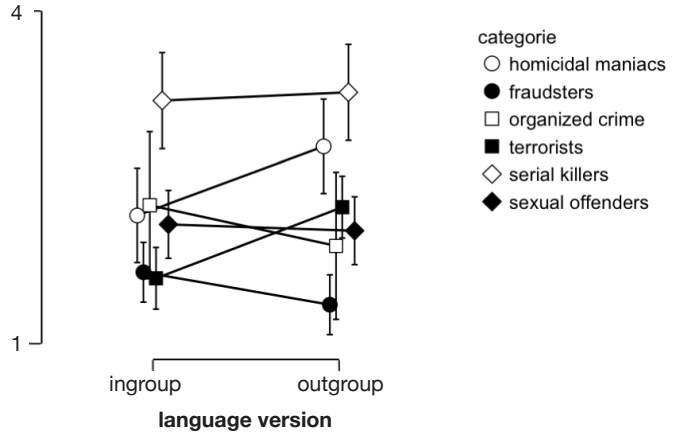

Figure 5. *Percentage of anger-related words as a function of language version of the Wikipedia article and perpetrator category*

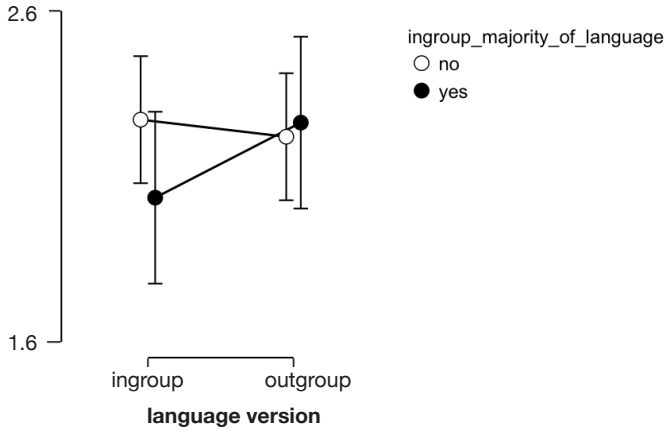

Figure 6. *Percentage of anger-related words as a function of language version of the Wikipedia article and whether the perpetrator in-group also constitutes the majority of authors contributing to that language version*

Categories of perpetrators. Since we had a substantial number of cases per category (see above), we explored in a Bayesian mixed-measures ANOVA with language version (in-group, out-group) as within-person variable and category (homicidal maniacs, fraudsters, organized crime, terrorists, serial killers, sexual offenders) as between-person variable whether the two factors might interact. As can be seen in in the BF_{10} column in Table 1, all models (except for the one with the main effect of language version only) received support from the data. Most support – also in the change from the prior model odds to the posterior model odds (BF_M) – received the model with the main effect of category only. Still, there was some weak support for the model including the interaction (see also the $BF_{inclusion}$ terms in Table 2). As displayed in Figure 5, there were virtually no differences between language versions for serial killers and sexual offenders, $BF_{01} = 4.61$ and $BF_{01} = 3.88$, respectively. But even for perpetrators from organized crime and fraudsters, who both showed a descriptively higher percentage of negative words in the in-group language version than the out-group language version, there was still more support for the null-hypothesis, $BF_{01} = 3.33$ and $BF_{01} = 1.74$, respectively. For homicidal maniacs and for terrorists, however, an alternative to the null-hypothesis was more likely given the data and while the evidence was weak for homicidal maniacs, $BF_{01} = 0.73$, it was stronger for terrorists, $BF_{01} = 0.07$, indicating that an alternative to the null-hypothesis was 15.33 times more likely given the data. And an inspection of the descriptive statistics shows that this alternative was contrary to the BSE-hypothesis: the percentage of negative emotion words was *lower* in the in-group language version compared to the out-group language version.

Heterogeneity of authorship. One major reason to suppose that Wikipedia is less prone to the BSE was the heterogeneity of the authors. Given that some languages are spoken by many people, others by fewer, the language versions have a differential potential to attract heterogeneity concerning the nationality of the authors: Wikipedia versions of languages that are spoken by a larger amount of people, such as English, Spanish and Arab are

Table 1. *Model-based analysis of the Bayesian mixed-measures ANOVA of percentage of negative emotion words with language version and perpetrator category*

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.200	8.743e-10	3.497e-9	1.000	
language version	0.200	1.845e-10	7.382e-10	0.211	2.490
category	0.200	0.603	6.088	6.902e+8	0.822
language version + category	0.200	0.128	0.585	1.460e+8	2.147
language version + category + language version ✻ category	0.200	0.269	1.471	3.075e+8	1.787

Note. All models include subject.

Table 2. *Analysis of effects, averaging across the models that contain a specific factor from the Bayesian mixed-measures ANOVA of percentage of negative emotion words with language version and perpetrator category*

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$
language version	0.600	0.397	0.438
category	0.600	1.000	6.296e+8
language version ✻ category	0.200	0.269	1.471

likely edited by a more heterogeneous group of authors than languages spoken by fewer people, such as Italian or Finnish[7]. In order to explore whether author heterogeneity affects the (potentially biased) presentation of perpetrators in Wikipedia articles, we reanalyzed our data. As a proxy for heterogeneity, we compared perpetrators, who had the same nationality as the majority of users generally authoring the Wikipedia language version (> 50% of the authors[7]; i.e., homogeneous language in-group version) with perpetrators who had the same nationality as a minority of users generally authoring that language version of Wikipedia (e.g., Portuguese perpetrators, as the Portuguese Wikipedia is mainly authored by Brazilians[7], but also any English- and Spanish-speaking perpetrators as none of the groups editing those language versions pose more than 50% of the authors).

A Bayesian mixed-measures ANOVA with language version (in-group, out-group) as within-person variable and perpetrator group (homogeneous, heterogeneous) as between-person variable and the percentage of negative emotion words as dependent variable yielded not much support for any of the factors or interactions as can be seen in Tables 3 and 4. In fact, the null-model received the greatest support from the data. Therefore, we did not gather support for the notion that an increased likelihood of in-group members as Wikipedia authors for the in-group article version had a substantial impact on our findings. Moreover, Figure 6 rather suggests that more in-group members likely contributing to the in-group language version rather led to a descriptive results pattern contrary to the BSE hypothesis.

7 https://stats.wikimedia.org/wikimedia/squids/SquidReportPageEditsPerLanguageBreakdown.htm [August 17, 2017]

Table 3. *Model-based analysis of the Bayesian mixed-measures ANOVA of percentage of negative emotion words with language version and ingroup majority (homogeneity)*

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Null model (incl. subject)	0.200	0.691	8.936	1.000	
language version	0.200	0.120	0.544	0.173	1.115
ingroup_majority_of_language	0.200	0.152	0.720	0.221	3.412
language version + ingroup_majority_of_language	0.200	0.028	0.117	0.041	8.048
language version + ingroup_majority_of_language + language version ✻ ingroup_majority_of_language	0.200	0.008	0.034	0.012	1.943

Note. All models include subject.

Table 4. *Model-based analysis of the Bayesian mixed-measures ANOVA of percentage of negative emotion words with language version and ingroup majority (homogeneity)*

Effects	P(incl)	P(incl|data)	$BF_{Inclusion}$
language version	0.600	0.157	0.124
ingroup_majority_of_language	0.600	0.189	0.156
language version ✻ ingroup_majority_of_language	0.200	0.008	0.034

Discussion

In this study, we set out to test whether the BSE occurs in Wikipedia – despite its norms and its heterogeneous authorship. By means of automatic text analyses and the percentage of negative emotion words in general or anger-related words specifically, we found no support at all for this hypothesis. Partly we found even strong support *against* it. Descriptively, the pattern was into the contrary direction with somewhat higher percentages of negative emotion words and anger-related words in the in-group language version compared to the out-group language versions. The null-hypothesis, that there is no BSE in Wikipedia, for which we had explicitly provided arguments in the introduction, received much more support from our data. Importantly, this support for the null-hypothesis cannot result from translations of articles between language versions. Not only is there research documenting that each language version is organically grown, that is, constructed uniquely by authors proficient of that language rather than being translated by some other language version (Hecht & Gergle, 2010; Stvilia, Al-Faraj, & Yi, 2009). Also, the profound difference in article length between ingroup and outgroup language version obtained in our study (and see also Hecht & Gergle, 2009; 2010) clearly speaks against this possibility.

The Black Sheep Effect – not an effect in Wikipedia?

The present study is the first to assess the BSE in a natural context with collaboratively written texts that are accessable to an imense amount of people. Most previous studies on the BSE presented negative behavior of a previously unknown target person, that conducted mild forms of negative behavior, such as providing a poor speech (e.g., Marques et al., 1988) and assessed participants' reception and their personal evaluation of the target directly (Otten & Gordijn, 2014). The present study is the only one we are aware of that examined the BSE in the *production* of presentations. Moreover, it is the first to investigate the BSE in

a context of *joint* production – i.e., collaboration – among a group of diverse authors and under the influence of norms that strongly promote unbiased presentations. And finally, it is the first to examine the BSE by making use of automatic text analyses (but see Greving et al., 2018, for an application of LIWC in the context of emotional contents of Wikipedia articles).

The fact that we did not obtain a more negative representation of perpetrators in the in-group articles when compared to the outgroup articles is particularly noteworthy in consideration of the fact, that we are talking about people who conducted severe negative behavior, such as killing others. Such severe misconducts should be more threatening and thus should elicit a more pronounced dissociation from these individuals when they are part of the in-group – in other words: produce a stronger BSE. Not only did our data not support this hypothesis at all but provided support for the null-hypothesis instead. The pattern of results was descriptively even contrary to the BSE-hypothesis. And in one instance – terrorists – we even found substantial support for such a *reverse* BSE effect, that is, a *less* negative presentation of terrorists in the ingroup language article version compared to the outgroup language article version. This is surprising and we can only speculate about the reasons. We have selected the target persons according to the pre-conditions of the BSE (Otten & Gordijn, 2014). We had no control over the integration of the perpetrator into the group, however, nor over the identification of authors with their nationality. Moreover, one might argue that the situation is not an inter-group context that renders group membership salient and therefore does not elicit the BSE (Turner et al., 1987). These boundary conditions could, however, only eliminate or prevent bias, but not produce a *reverse* bias.

We are not aware of any lab study tackling terrorists or homicidal maniacs, let alone systematic comparisons of different types of misconduct. Possibly, the surprising results are due to subtyping in extreme cases of negative behavior. Subtyping is the process where group members that disconfirm the prototype of the group are labeled as "exceptions", so that the group prototype remains unchanged (Maurer, Park, & Rothbarth, 1995). Extreme cases of deviance are more easily subtyped than moderate cases (Johnston & Hewstone, 1992). Thus, our targets may have been subtyped in the authors' minds. It remains unclear, however, why perpetrators were still sometimes evaluated more positively by the in-group than the out-group. Possibly, the perpetrators are put into a new category of "deviants", where they no longer constitute a threat to the in-group image. Self-categorization can take place on several layers of multiple group memberships (Turner et al., 1987), and there is evidence that others' belonging to multiple social categories is applied strategically as a self-management technique (Stelzl, Janes, & Seligman, 2008). Once perpetrators are put into the "extra" category, the common nationality (with some authors) might be a side-aspect of the person that constitutes a similarity between authors and perpetrator. This similarity might, even though subordinated, open the gates for a positive bias. This is, however, speculation. After all, the pattern of results does not allow for a differentiation between a positive bias regarding the ingroup (i.e., ingroup favoritism) and a negative bias regarding the outgroup (i.e., outgroup derogation; e.g., Hewstone et al., 2002; March & Graham, 2015; Olsson, Ebert, Banaji, & Phelps, 2005). Moreover, it must be acknowledged, that we had found only weak support for an interaction between perpetrator category and article version and thus – overall – much more support for the null-hypothesis. But it might inspire future research to take a look at different perpetrator categories or to compare moderate and extreme negative behaviors. Also, the perception of perpetrators in terms of multiple social categories seems to be a fruitful avenue for future research.

Potentials of socially negotiated knowledge representations

Wikipedia is a socially negotiated compendium of content that is characterized by two things: (1) the norms of neutrality and verifiability and (2) the potential heterogeneity of the authorship. Our results might suggest that both characteristics could have contributed to the elimination of biases over a range of perpetrator categories. The overall support for the null-hypothesis might be due to the norms promoting unbiased information and in line with research that demonstrates the impact of such norms (e.g., Postmes et al., 2001). Prior research on the hindsight bias – a very robust and pervasive error in human judgment (Guilbault et al., 2004; Roese & Vohs, 2012) – has shown, for instance, that it was much less frequently found in Wikipedia articles than we would expect from research on individuals' subjective perceptions and evaluations. Only for severe negative events – disasters – was there evidence for a hindsight bias in Wikipedia (Oeberst et al., 2018). Moreover, since Wikipedia exists in language versions and not country-versions, authorship is more or less diverse. The present data hints towards the notion that the heterogeneity of author nationality bears the potential to reduce biases in Wikipedia: The explorative comparison of perpetrators who do or do not belong to the nationality of the likely majority of authors of that language version of Wikipedia did not show a significant difference, but the descriptive pattern was even more in line with the null-hypothesis for the case with a likely more heterogeneous authorship. Although one should always be careful when interpreting null-effects, the findings are in line with the diversity literature (e.g., van Knippenberg, de Dreu, & Hohman, 2004) and literature on decision making processes (Schulz-Hardt, et al., 2000; Vinokur & Burnstein, 1978), which documented a reduction or even elimination of bias due to heterogeneity.

Limitations and future prospects

The ecological validity of the material is the charm and the curse of the present findings. It is, for instance, difficult to track the actual nationalities of the authors[8] (which we derive from language or average statistics), so that one might criticize that the categorization into in-group and out-group articles may be faulty. The fact that we found systematic evidence for an in-group bias in articles about inter-group conflicts, however, clearly argues against this notion (Oeberst, von der Beck, Matschke, Ihme, & Cress, 2019; Oeberst, Ihme, Matschke, & Cress, 2019). Similarly, we cannot clearly state that the adherence of the authors to the norms of Wikipedia is the ultimate process reducing biases in articles because we neither measured nor manipulated them. Future research should therefore complement the present findings by replicating them under controlled laboratory settings.

One might also criticize the use of automatic text analyses as main dependent variable, because it only offers limited insight into the content. Thus, it does not allow the investigation of other, more subtle forms of inter-group bias (e.g., Hewstone, 1990; Maass et al., 1989; Oeberst & Matschke, 2017). The inclusion of very different languages, however, poses a challenge. An elaborate content analysis, for instance, would either need translated materials, which is not only cost-intensive, but has its own drawbacks (e.g., Winter,

8 Users may not only contribute anonymously to Wikipedia but also often do not self-identify their nationality on their user-page when registered.

2007), or coders, who were fluent in very many languages, which is unrealistic. Automatic text analyses enable, in contrast, the analysis of original articles (rather than translated versions), and are objective and economic and have proven valid in previous research (see Boyd & Pennebaker, 2015; Tausczick & Pennebaker, 2010). Moreover, LIWC has been successfully applied to analyze emotional contents in Wikipedia articles previously (Greving et al., 2018). Nevertheless, future research should aim at replicating the findings with other methods (e.g., content coding) and other dependent measures.

References

Boyd, R. L., & Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. In C. Dimofte, C. Haugtvedt, & R. Yalch (Eds.), *Consumer psychology in a social media world* (pp. 222-236). New York City, NY: Routledge Publishers.

Braun, M. (2010). Dealing with a deviant group member. Kurt Lewin Institute Dissertation Series, 2010-12. Enschede, NL: GildePrint.

Braun, M., Otten, S., & Gordijn, E. H. (2009). Did you really mean that? The role of group affiliations and ambiguity of negative intentions in aggressive interactions. Unpublished manuscript, University of Groningen, The Netherlands.

Callahan, E. S., & Herring, S. C. (2011). Cultural bias in Wikipedia content on famous persons. *Journal of the American Society for Information Science and Technology, 62,* 1899-1915. doi:10.1002/asi.21577

Fast, L. A, & Funder, D. C. (2008). Personality as manifest in word use: Correlations with self-report, acquaintance report, and behavior. *Journal of Personality and Social Psychology, 94,* 334-346. doi:10.1037/0022-3514.94.2.334

Fehr, J., Sassenberg, K., & Jonas, K. J. (2012). Willful stereotype control: The impact of internal motivation to respond without prejudice on the regulation of activated stereotypes. *Zeitschrift für Psychologie – Journal of Psychology, 220(3)*, 180-186. doi:10.1027/2151-2604/a000111

Forte, A., & Bruckman, A. (2008). Scaling Consensus: increasing decentralization in Wikipedia governance. *Proceedings of the International Conference of the Learning Sciences, Bloomington, USA.*

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438,* 900-901. doi:10.1038/438900a

Greving, H., Oeberst, A., Kimmerle, J., & Cress, U. (2018). Emotional content in Wikipedia articles on negative man-made and nature-made events. *Journal of Language and Social Psychology, 37,* 267-287. doi:10.1177/0261927X17717568

Guilbault, R. L., Bryant, F. B., Brockway, J. H., & Posavac, E. J. (2004). A meta-analysis of research on hindsight bias. *Basic and Applied Social Psychology, 26,* 103–117. doi:10.1080/01973533.2004.9646399

Hecht, B., & Gergle, D. (2009). Measuring self-focus bias in community-maintained knowledge repositories. *Proceedings of the 4th International Conference on Communities and Technologies,* pp. 11-20. doi:10.1145/1556460.1556463

Hecht, B., & Gergle, D. (2010). The tower of Babel meets web 2.0: user-generated content and its applications in a multilingual context. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 291-300).

Hewstone, M. (1990). The 'ultimate attribution error'? A review of the literature on intergroup causal attribution. *European Journal of Social Psychology, 20,* 311-335. doi:10.1002/ejsp.2420200404

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes Factors. *Journal of Problem Solving, 7,* 2-9. doi:10.7771/1932-6246.1167

Johnston, L., & Hewstone, M. (1992). Cognitive Models of stereotype change 3. Subtyping and the perceived typicality of disconfirming group members. *Journal of Experimental Social Psychology, 28,* 360-386. doi:10.1016/0022-1031(92)90051-K

Maass, A., Salvi, D., Arcuri L., & Semin, G. R. (1989). Language use in intergroup context. *Journal of Personality and Social Psychology, 57,* 981-993. doi:10.1037//0022-3514.57.6.981

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: thinking categorically about others. *Annual Review of Psychology, 51,* 93-120. doi:10.1146/annurev.psych.51.1.93

March, D. S., & Graham, R. (2015). Exploring implicit ingroup and outgroup bias toward Hispanics. *Group Processes & Intergroup Relations, 18,* 89-103. doi:10.1177/1368430214542256

Marques, J. M., & Yzerbyt, V. Y. (1988). The black sheep effect: Judgmental extremity towards ingroup members in inter- and intra-group situations. *European Journal of Social Psychology, 18,* 287-292. doi:10.1002/ejsp.2420180308

Marques, J., Abrams, D., & Serôdio, R. G. (2001). Being better by being right: Subjective group dynamics and derogation of in-group deviants when generic norms are undermined. *Journal of Personality and Social Psychology, 81,* 436-447. doi:10.1037/0022-3514.81.3.436

Marques, J. M., Yzerbyt, V. Y., & Leyens, J. (1988). The 'Black Sheep Effect': Extremity of judgments towards ingroup members as a function of group identification. *European Journal of Social Psychology, 18,* 1-16. doi:10.1002/ejsp.2420180102

Maurer, K. L., Park, B., & Rothbarth, M. (1995). Subtyping versus subgrouping processes in sterotype representation. *Journal of Personality and Social Psychology, 69,* 812-824.

Monteith, M. J., Sherman, J. W., & Devine, P. G. (1988). Suppression as a stereotype control strategy. *Personality and Social Psychology Review, 2,* 63-82. doi:10.1207/s15327957pspr0201_4

Oeberst, A., Halatchliyski, I., Kimmerle, J., & Cress, U. (2014). Knowledge construction in Wikipedia: A systemic-constructivist analysis. *Journal of the Learning Sciences, 23,* 149-176. doi:10.1080/10508406.2014.888352

Oeberst, A., Ihme, T. A., Matschke, C., & Cress, U. (2019). Picture us and them: Inter-group Bias in pictures of Wikipedia articles about inter-group conflicts. *Manuscript submitted for publication.*

Oeberst, A., & Matschke, C. (2017). Word order and world order. Titles of intergroup conflicts may increase ethnocentrism by mentioning the in-group first. *Journal of Experimental Psychology: General, 146,* 672-690. doi:10.1037/xge0000300

Oeberst, A., von der Beck, I., Back, M., Cress, U., & Nestler, S. (2018). Biases in the production and reception of collective knowledge: The case of hindsight bias in Wikipedia. *Psychological Research, 82,*1010-1026.. doi:10.1007/s00426-017-0865-7

Oeberst, A., von der Beck, I., Matschke, C., Ihme, T. A., & Cress, U. (2019). Divergent perspectives on history: In-group Bias in Wikipedia articles about inter-group conflicts. *Manuscript under revision.*

Olsson, A., Ebert, J., P., Banaji, M. R., & Phelps, E. A. (2005). The role of social groups in the persistence of learned fear. *Science, 309,* 785-787. doi:10.1126/science.1113551

Otten, S., Gordijn, E. H. (2014). Was it one of us? How people cope with misconduct by fellow ingroup members. *Social and Personality Psychology Compass, 8,* 165-177. doi:10.1111/spc3.12098

Pinto, I. R., Marques, J. M., Levine, J. M., & Abrams, D. (2010). Membership status and subjective group dynamics: Who triggers the black sheep effect? *Journal of Personality and Social Psychology, 99,* 107-119. doi:10.1037/a0018187

Postmes, T., Spears, R., & Cihangir, S. (2001). Quality of decision making and group norms. *Journal of Personality and Social Psychology, 80,* 918-930. doi:10.1037/0022-3514.80.6.918

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives in Psychological Science, 7,* 411-426. doi:10.1177/1745691612454303

Schultheiss, O. C. (2013). Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. *Frontiers in Psychology, 4,* Article 748. doi:10.3389/fpsyg.2013.00748

Schulz-Hardt, S., Frey, D., Lüthgens, C., & Moscovici, S. (2000). Biased information search in group decision making. *Journal of Personality and Social Psychology, 78,* 655-669. doi:10.1037//0022-3514.78.4.655

Stelzl, M., Janes, L., & Seligman, C. (2008). Champ or chump: stratigic utilization of dual social identities of others. *European Journal of Social Psychology, 38,* 128-138. doi:10.1002/ejsp.446

Stvilia, B., Al-Faraj, A., & Yi, Y. (2009). Issues of cross-contextual information quality evaluation — The case of Arabic, English, and Korean Wikipedias. *Library & Information Research, 31,* 232-239. doi:10.1016/j.lisr.2009.07.005

Tajfel, H., & Turner, J. (1979). An integrative theory of inter-group conflict. In J. A. Williams & S. Worchel (Eds.), *The social psychology of inter-group relations* (pp. 33-47). Belmont, CA: Wadsworth.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized Text Analysis Methods. *Journal of Language and Social Psychology, 29,* 24-54. doi:10.1177/0261927X09351676

Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). *Rediscovering the social group. A self-categorization theory.* Oxford: Basil Blackwell.

van Knippenberg, D., de Dreu, C. K. W., & Hohman, A. C. (2004). Work group diversity and group performance: An integrative model and research agenda. *Journal of Applied Psychology, 89,* 1008-1022. doi:10.1037/0021-9010.89.6.1008

van Leeuwen, E., ban den Bosch, M., Castano, E., & Hopman, P. (2010). Dealing with deviants: The effectiveness of rejection, denial, and apologies on protecting the public image of a group. *European Journal of Social Psychology, 40,* 282-299. doi:10.1002/ejsp662

van Prooijen, J. (2006). Retributive reactions to suspected offenders: The importance of social categorizations and guilt probability. *Personality and Social Psychology Bulletin, 32,* 715-726. doi:10.1177/0146167205284964

Viégas, F. B., Wattenberg, M., Kriss, J., & van Ham, F. (2007). Talk Before You Type: Coordination in Wikipedia. *Proceedings of the 40th Annual Hawaii International Conference on System Sciences.*

Vinokur, A., & Burnstein, E. (1978). Depolarization of attitudes in groups. *Journal of Personality and Social Psychology, 36,* 872-885. doi:10.107/0022-3514.36.8.872

Winter, D. G. (2007). The role of motivation, responsibility, and integrative complexity in crisis escalation: comparative studies of war and peace crises. *Journal of Personality and Social Psychology, 92,* 920-937. doi:10.1037/0022-3514.92.5.920

Wolf, M., Sedway, J., Bulik, C. M., & Kordy, H. (2007). Linguistic analyses of natural written language: unobstrusive assessment of cognitive style in eating disorders. *International Journal of Eating Disorders, 40,* 711-717. doi:10.1002/eat

Received: 7.27.2018
Revised: 9.10.2018
Accepted: 9.11.2018