Episteme (2022), 1–35
doi:10.1017/epi.2022.21

ARTICLE

Epistemic Side-Effect Effect: A Meta-Analysis

Bartosz Maćkiewicz, Katarzyna Kuś, Katarzyna Paprzycka-Hausman and Marta Zaręba

Faculty of Philosophy, University of Warsaw, Warsaw, Poland
Corresponding author: Bartosz Maćkiewicz. Email: b.mackiewicz@uw.edu.pl

(Received 1 December 2021; revised 18 March 2022; accepted 18 May 2022)

Abstract

Beebe and Buckwalter (2010) made the surprising discovery that people are more inclined to attribute knowledge when norms are violated than when they are conformed to. The epistemic side-effect effect (ESEE) is the analogue of the Knobe effect (Knobe 2003a). ESEE was replicated in a number of experiments. It was also studied under various conditions. We have carried out a meta-analysis of research on ESEE. The results suggest that ESEE is a robust finding but its magnitude is highly variable. Two study-level covariates influence its size: the subject of the knowledge attribution (agent vs third-party) and the type of norm that is violated or complied with. The effect size is not influenced, however, by the manipulation of chances, by whether the story is about a side effect or not, by language or by question phrasing. The impact of the Gettierization of the story is marginally significant.

Keywords: Epistemic side-effect effect; meta-analysis; knowledge ascription; experimental philosophy

Introduction

The epistemic side-effect effect (ESEE) is an epistemic counterpart of the Knobe effect. Knobe (2003*a*) discovered that people tend to ascribe intentionality to unintended side effects of an intentional action when the side effect violates some norm (Knobe 2007, 2010; Holton 2010; Robinson *et al.* 2015) but not when it does not. Although the Knobe effect was foreshadowed in philosophical discussions (Harman 1976), its robustness was of concern especially to theorists dealing with intentional action. Beebe and Buckwalter (2010) found a similar effect for knowledge attributions. They used the environment vignettes of Knobe's (2003*a*) study but asked participants about knowledge rather than intentionality. Quite unexpectedly, they found that the attributions of knowledge that a norm-violating side effect would occur are stronger than the attributions of knowledge that a norm-conforming side effect would occur. These results, which have been repeated in many later studies, pose a challenge to traditional epistemology, according to which the normative valence of a result has no bearing on the attribution of knowledge.

The main aim of this paper is to assess the existence and the magnitude of ESEE by means of meta-analysis. We have identified eight study-level factors, which correspond

© The Author(s), 2022. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

to the most common types of design changes with respect to Beebe and Buckwalter's (2010) experiment (section 1). In section 2, we describe the methods we have used in data selection and analyses. In section 3, we present the results of various meta-analyses conducted. The results are then discussed in section 4.

1. Study-level factors

Most studies on ESEE introduce some variations to the experimental design by Beebe and Buckwalter (2010). Only a small minority of studies are direct replications of the original experiment. The most common types of such changes include: the type of norm involved, the type of scale used, the phrasing of the question, the language of the study, the manipulation of the chances of the side effect, the Gettierization of the story, the type of effect involved, and the subject of the knowledge attribution. Using a meta-analytical approach, we wanted to check whether the asymmetry in knowledge attributions in these cases is comparable in magnitude to the original effect size found by Beebe and Buckwalter (2010). A summary of all study-level factors alongside with the coding scheme that was used in the meta-analyses can be found in Table 2.

1.1. Type of norm

The study conducted by Beebe and Buckwalter (2010) used the environment vignettes introduced by Knobe (2003*a*). In brief, the chairman of a company is told that a new programme will increase the company's profits but (and) will also have the side effect of harming (helping) the environment; the chairman does introduce the programme thus harming (helping) the environment.

There is a consensus in the literature that the norm ("do not harm the environment") that is violated in the negative condition has a moral (or a quasi-moral) character. The Knobe effect was successfully replicated for other kinds of norms: aesthetic (Knobe 2004), legal (Knobe 2007) and conventional (Knobe and Mendlow 2004) norms. As an epistemic counterpart to the Knobe effect, ESEE was also tested on a variety of stories where other kinds of norms were violated (Beebe and Jensen 2012; Beebe 2016). It is natural to suppose that violations of moral norms generate a more pronounced asymmetry compared with non-moral ones. Indeed, moral considerations have been shown to affect the attribution of various mental and non-mental concepts (for a summary and discussion, see e.g. Knobe *et al.* 2012). We thus decided that we will focus on whether the norm is moral or not.

We encountered two problems when coding the studies. First, more than one norm was invoked in some vignettes. For example, in Knobe's (2007) Nazi study, there is a moral but also a legal norm. An immoral law is violated in the negative condition, while it is complied with in the positive condition. Second, the question that was asked of the participants could make one norm more salient. Indeed, in the Nazi study, the question that is asked concerns the violation of the legal (rather than the moral) norm. In Nadelhoffer's (2004) dice scenarios, some questions concerned killing (violation of a moral norm) while others concerned throwing a six (there is no general norm against throwing a six).

To take these problems into account, we decided to code the studies in three different ways:

- *Present* (Is any moral norm present?) adopts the value "true" if some moral norm is in play in the negative condition; it adopts the value "false" otherwise;

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

- *Salient* (Is any moral norm salient?) adopts the value "true" if some moral norm is contextually salient in the negative condition; it adopts the value "false" otherwise (if the salient norm is not moral or if no norm is made salient);
- *Violated* (Is any moral norm violated?) – adopts the value "true" if some moral norm is violated (directly or indirectly) by the agent's action in the negative condition; it adopts the value "false" otherwise.

It should be stressed that all coding schemes were based only on the negative condition. This was to ensure a consistent coding especially in cases where multiple norms were at stake. We took the norm to be contextually salient when its content was directly involved in the content of the knowledge claim. For example, knowledge that the environment would be harmed or that someone would be killed was taken to involve a salient moral norm. By contrast, knowledge that organizational changes would violate the requirements of the law was taken to involve a salient non-moral (legal) norm (*Salient* was assigned the value "false"). Likewise knowledge that Donald is in Italy or that the die comes up six was not taken to involve any salient norm (*Salient* was assigned the value "false").

To illustrate the difference between these three coding schemes (see Table 1), let us consider the Nazi study (Knobe 2007; Beebe and Jensen 2012) and the Dice study (Nadelhoffer 2004; Paprzycka-Hausman 2020). In the Nazi study, a legal norm rather than a moral norm is contextually salient since the participants were asked whether "the CEO knew that the organizational changes would violate the requirements of the law" (Beebe and Jensen 2012). However, a moral norm is also in play – complying with the law would result in some people being sent to concentration camps, while violating the law might save them from the horrible fate (*Present* adopts the value "true"). The Nazi study is structured in such a way that the violation of the salient (non-moral) norm (in the negative condition) involves complying with the moral norm. In other words, in the negative condition, on which the coding decisions are based, no moral norm is violated (*Violated* adopts the value "false").

In the negative condition of Nadelhoffer's (2004) dice scenarios (Immoral Dice), throwing a six results in detonating a bomb and killing Smith. It is thus rather clear that a moral norm is violated and thus that some moral norm is present (*Present* and *Violated* are thus "true"). Whether a moral norm is salient depends on the question asked. Indeed, in one group (Immoral Dice – Killing), participants were asked about the intentionality of killing Smith. Clearly, the norm that is made salient by this question is moral. In another group, which was given the very same vignette, the question asked was about the intentionality of throwing a six (Immoral Dice – Six). In the latter group, it is not so clear whether the norm that is made salient by the question is moral. On a narrow interpretation, since there is no norm against throwing dice (including throwing six) in general, the question does not make any moral norm salient. On a wider interpretation, one could argue that the question does make a moral norm salient since, in this particular case, the protagonist *ought not to* throw the die at all given the possible consequence of throwing a six.

We decided to adopt the narrow interpretation of *Salient* for at least two reasons. First, on the wider interpretation, there would be no difference between the *Salient* and the *Violated* coding schemes. Second, Nadelhoffer's (2004) study shows that people's responses are affected by what norm is salient in norm-violation conditions, which justifies the introduction of both coding schemes. In the Immoral Dice scenario, he obtained a striking difference in the attributions of intentionality to killing Smith (87.5%) and

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Table 1. The type norm in the Nazi study (Knobe 2007; Beebe and Jensen 2012) and in Dice studies (Nadelhoffer 2004; Paprzycka-Hausman 2020; see text for explanation) according to three coding schemes

	Nazi	Dice – killing	Dice – throwing six
Present	+	+	+
Salient	–	+	–
Violated	–	+	+

throwing a six (55%). In this scenario (irrespective of the question that is asked), Smith is killed and thus a moral norm is violated (*Violated* = true in both groups "Immoral Dice – Killing" and "Immoral Dice – Six"). The *Salient* coding scheme allows to capture the difference between the group that is asked about killing Smith (*Salient* = true in the "Immoral Dice – Killing" group) and the group that is asked about throwing a six (*Salient* = false in the "Immoral Dice – Six" group). A moral norm is salient only when participants are explicitly asked whether Smith was killed intentionally.

1.2. Scale

In contrast to Knobe's (2003*a*) study on the attribution of intentionality, Beebe and Buckwalter (2010) used a non-dichotomous scale. Beebe and Jensen (2012) were the first to raise the problem that the specifics of the scale used to collect responses may matter. However, they attempted to show that ESEE occurs regardless of the scale used. In the literature on ESEE, several types of non-dichotomous scales are present:

- 5-point Likert scale (Turri 2014; Paprzycka-Hausman 2021*c*),
- 7-point Likert scale from −3 to 3 (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Buckwalter 2014; Beebe 2016; Paprzycka-Hausman 2020),
- 7-point Likert scale from 1 to 7 (Beebe and Jensen 2012; Beebe and Shea 2013; Beebe 2016; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021),
- 8-point Likert scale (Ziółkowski *et al.* Ms *a*, *b*),
- 100-point visual analogue scale (Dalbauer and Hergovich 2013).

Relatively few studies have been conducted on dichotomous scales (e.g. Beebe and Jensen 2012; Paprzycka-Hausman 2020). Finally, some studies use so-called weighted scales, which combine binary responses with confidence ratings (Turri 2014; Yuan and Kim 2021). They constitute an interesting middle ground between dichotomous and non-dichotomous scales. The weighting makes them comparable to non-dichotomous scales but the fact that participants need to take a clear stand on the question makes them comparable to dichotomous scales. We coded them as non-dichotomous scales.

1.3. Question phrasing

Two different question probes are used across research on ESEE. In several studies, participants were asked directly about the protagonist's knowledge (knowledge probe). For example, they were asked "Did the chairman know that the program would harm/help the environment?". Such a probe was used, among others, by Beebe and Buckwalter (2010), Beebe (2013) and Turri (2014). In other studies, participants were asked to

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

what extent they agree or disagree with a claim about the protagonist's knowledge (agreement probe). For example, they were asked "To what extent do you agree that the chairman knew that the program would harm/help the environment?". The agreement probe was used for example by Dalbauer and Hergovich (2013), Buckwalter (2014) and Paprzycka-Hausman (2020).

Interestingly, this distinction in the phrasing of the question does not overlap with the use of dichotomous and non-dichotomous scales. Several studies (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013) used a knowledge probe and a non-dichotomous scale. On the other hand, Yuan and Kim (2021) used an agreement probe and a dichotomous scale (as their first question). Next, they inquired about the level of confidence on a non-dichotomous scale.

Questions differed also with regard to the exact content of the knowledge claim. We address this issue separately (see section 1.9).

1.4. Language

The vast majority of research on ESEE has been conducted in English. However, ESEE has been replicated in German (Dalbauer and Hergovich 2013), Mandarin Chinese (Yuan and Kim 2021), Korean (Yuan and Kim 2021), and Polish (Ryszkowska *et al.* Ms; Zaręba Ms). However, very few of those studies were replications of Beebe and Buckwalter's (2010) experiment. Only the German study and one of the Polish studies (Zaręba Ms) employed a translated version of Knobe's (2003*a*) environment story. Moreover, the German study used a 100-point visual analogue scale and an agreement probe rather than the knowledge probe.

In the analysis, we decided to consider whether ESEE is in some way specific to English speakers or more pronounced in English. All non-English studies were thus coded as "other".

1.5. Chances

Beebe and Buckwalter (2010) did not manipulate the probability of the side effect. They used Knobe's (2003) environment vignettes where the chances are not specified and open to participants' interpretation. Some researchers hypothesized that at least some part of the effect can be explained by the fact that participants' estimations of probabilities differ in the positive cases and in the negative cases. Indeed, chances were manipulated in several experiments (e.g. Beebe and Jensen 2012; Paprzycka-Hausman 2020). In all studies where probabilities were explicitly stated to be low ("low chances", "slight chances", "1%", etc.), *Chances* adopts the value "low". It adopts the value "normal" otherwise, i.e. for experiments where either the chances were stated to be high ("high chances", "99%", "certain", "very strong chance" etc.) or they were not stated at all. Notably, the first study on ESEE that manipulated chances (Experiment 5 in Beebe and Jensen 2012) is not included in the analysis because the authors' manipulation was not systematic enough. There were only two groups in the experiment: the low-chance harm condition and the high-chance help condition (see section 2.2).

1.6. Gettierization

Several studies (Beebe and Shea 2013; Buckwalter 2014; Turri 2014; Yuan and Kim 2021) concerned the Gettierized epistemic side-effect effect (GESEE). In Gettierized cases, the protagonist has a true and justified belief but some epistemic luck is involved.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Indeed, research shows that people tend not to attribute knowledge in Gettierized cases (Beebe and Shea 2013). However, respondents are more willing to attribute knowledge in Gettierized Knobe-type cases where the outcome of the action is negative. GESEE was tested both on Gettierized Knobe-type scenarios (Beebe and Shea 2013) as well as on new vignettes (Beebe and Shea 2013; Buckwalter 2014; Turri 2014; Yuan and Kim 2021). The results obtained for English have been replicated to some extent in Korean and Mandarin Chinese (Yuan and Kim 2021).

1.7. Effect type

The vast majority of the studies concerned the asymmetrical attribution of knowledge in scenarios that involved a side effect. However, an asymmetry in knowledge attributions between normatively negative and positive cases was also found in scenarios where there is no side-effect effect: some Gettier-type scenarios (e.g. Beebe and Shea 2013) or Butler-like scenarios (e.g. Paprzycka-Hausman 2020). We decided to include those studies as well but to distinguish them from the others by taking the effect type into consideration.

This decision might look controversial in view of the fact that the very name of the effect under consideration mentions side effects. Indeed, some authors take the Knobe effect to be the discovery that people tend to attribute intentionality in cases where the side effect of an intentional action violates a norm but not in cases where it conforms to a norm. It should be pointed out, however, that it is not uncontroversial to view the Knobe effect as the side-effect effect for intentionality attributions. After all, the asymmetry in intentionality attributions between normatively positive and negative cases was empirically discovered and intuitively predicted not only for the side-effect scenarios but also for Butler-type scenarios where the resulting action (main effect) is obtained via luck rather than skill (Harman 1976; Butler 1978; Knobe 2003*a*, *b*). Moreover, most accounts of the Knobe effect typically apply to both kinds of cases (e.g. Hindriks 2008, 2011; Knobe 2010; Alfano *et al.* 2012). It is, of course, historically true that ESEE has been found for Knobe's (2003*a*) side-effect scenarios. However, to the extent that a unified account of the intentionality attributions and knowledge attributions is at all conceivable (see e.g. Alfano *et al.* 2012), it is worth including both types of situations in which the original asymmetries in intentionality attributions were found.

1.8. Subject of the knowledge attribution

Beebe and Buckwalter (2010) asked the participants whether the protagonist of the story, the chairman, knew that the side effect would occur. In several other studies (Beebe and Shea 2013; Buckwalter 2014; Beebe 2016; Yuan and Kim 2021), the question asked concerned not the protagonist of the story but some other person (third-party observer, etc.). In some but not all of these studies, a similar asymmetry in knowledge attributions was observed.

It is worth pointing out that the replicability of ESEE for the third-party subject is not convincingly established. Some authors have found that the attribution of knowledge to the third person differs from the attributions of knowledge to the agent (Buckwalter 2014; Yuan and Kim 2021) while others have not (Beebe 2016). Meta-analysis should help in evaluating the strength of the evidence for the third-party ESEE and in comparing its magnitude to the first-party ESEE.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

1.9. Factors not taken into account in the meta-analysis

Let us briefly note several factors that differentiate the various versions of ESEE studies, which we decided not to consider in our meta-analysis.

The studies differ in how the side effect is described as well as in the content of the knowledge attribution. In some vignettes, it is just described generally as harming/helping the environment. In others, a more detailed description is given, e.g. as harming/improving water quality (Beebe and Shea 2013; Turri 2014) or as being poisonous/beneficial to the crops (Buckwalter 2014; Yuan and Kim 2021). Some researchers ask whether the agent knew that his or her action would have a certain (side) effect ("The mayor knew that *by signing the contract he would create/cut jobs*", see Buckwalter 2014). Others ask whether the agent knew that the result of his action would have a certain side effect (e.g. "The chairman knew that *the new program would harm/help the environment*", see for example Beebe and Buckwalter 2010; Beebe and Jensen 2012; Dalbauer and Hergovich 2013). Still others ask whether the agent knew that the side effect would occur (e.g. "The CEO knew that *local water quality levels were going to rise/fall*", see for example Beebe and Shea 2013). In many studies, it is unclear what the exact content of the knowledge claim is. We did not take such differences in content as a separate factor in the meta-analysis.

We did not check whether the asymmetry in knowledge attribution is affected by whether the agent is an individual or a group (the group effect was examined in Ziółkowski *et al.* Ms *a*, *b* and Ryszkowska *et al.* Ms). We also did not consider factors related to the method of conducting the study or recruitment procedure, for example whether the respondents were remunerated, whether they were college students, whether they were recruited through social media or through crowdsourcing websites such as mTurk or Prolific.

2. Materials and methods

2.1. Data sources and searches

A literature search for research on ESEE was performed by means of Google Scholar. First, a list of all papers that cite Beebe and Buckwalter's (2010) paper was generated. We assumed that papers that present new experimental data on ESEE would refer to the original experiment. The list was then analysed by two of us (KK and BM) to single out those studies that contain new results. Several studies that are not published at this moment were also included. For safety, meta-analysis was conducted for all studies as well as, independently, for only those studies that were published.

2.2. Study selection

We selected only studies that simultaneously satisfy the following three criteria. First, the design must include a negative and a positive (or neutral) condition. Second, the participants must be asked about knowledge. Third, the experiment must have a between-subject design.

Negative and positive/neutral condition. We included studies that involved various kinds of norms (moral, legal, aesthetic, etc.). However, we excluded experiments that used only a norm-violating or only a norm-conforming/neutral version of the vignette. A good example is the study reported by Beebe and Jensen (2012: 700), which manipulated the character of the agent but only in negative conditions (the norm-violating side effect was a person dying of starvation).

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

We also decided to include studies that had neutral rather than positive conditions. In some studies on Getterized ESEE (e.g. Beebe and Shea 2013), three conditions were present: a neutral condition ("unGettiered"), a Gettierized neutral condition ("Gettier") and a Gettierized harm condition ("Significant Harm"). We treated the latter two ("Gettier" and "Significant Harm") as a pair with "Significant Harm" as the negative condition.

Question about knowledge. Many phenomena are described as ESEE. There is a growing body of literature concerning ESEE for belief attribution (e.g. Alfano *et al.* 2012; Beebe 2013; Robinson *et al.* 2015) or justification attribution (Turri 2014). Our focus was on the attributions of knowledge. We have thus employed a simple rule: a study was selected if participants were asked about their attribution of knowledge to the protagonist or the observer. This rule allowed to include studies where other questions (e.g. questions about belief attribution) were *also* asked.

Between-subject design. We decided to include only between-subject studies. In general, one can doubt whether between-subject and within-subject experiments test the same phenomena (Charness *et al.* 2012; Ziółkowski 2017). In the literature on the Knobe effect, there are a few within-subject studies (see Nichols and Ulatowski 2007; Pinillos *et al.* 2011) but the same general worry can be raised. As a matter of fact, it turned out that there are no within-subject studies on ESEE yet.

2.3. Data extraction

The selection of studies was done by two of us (KK and BM). Disagreements were resolved in a group discussion with the remaining authors. Two of us performed the data extraction task. For each study, we identified the number of participants per group together with the descriptive statistics and we coded 8 study-level covariates. The resulting data were cross-checked twice by the remaining authors. In cases where the necessary statistics were not reported in a paper, they were supplemented by the authors.[1]

2.4. Data imputation and conversion

We used standardized mean difference (Cohen *d*, SMD) as a measure of the effect size. There were two problems with that choice. First, the majority of studies did not report this statistic. Second, for experiments with a forced-choice paradigm, the appropriate effect size must be computed and then converted to *d*. For data imputation (missing statistics) and conversion (from Odds Ratio to Cohen's *d*), we used the procedure and equations described in Appendix A.

2.5. Data synthesis and analysis

Two types of analyses were carried out. First, we assessed the overall meta-analytic evidence for the stability, size, and heterogeneity of ESEE. The analysis also allows to assess the publication bias. Second, meta-regression with selected study-level covariates was performed.

For factorial designs, i.e. for the majority of selected experiments, each pair of subgroups (negative vs positive/neutral subgroup) was treated as a unit of analysis (Borenstein *et al.* 2009), i.e. as a single study (henceforth we use the term 'study' in

[1]All researchers we contacted provided us with the additional data, for which we are very grateful.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

this sense). The resulting analysis is thus more granular with respect to selected covariates (section 3.6). In other words, an experiment that also manipulates factors other than the normative valence of the side effect was counted as multiple studies. For example, Beebe and Jensen's (2012) experiment 1 has a 2 (normative valence of the side effect: positive vs negative) × 2 (type of scale: dichotomous vs Likert) design. For the purpose of our analyses, it was counted as two studies.

Overall meta-analytic effect. Four analyses of the first type were conducted: (1) meta-analysis of all (published and unpublished) studies (section 3.2), (2) meta-analysis of all published studies (section 3.3), (3) meta-analysis of close replications (section 3.4) and (4) meta-analysis of studies on third-party ESEE (section 3.5). We took a close replication to involve studies that used Knobe's (2003*a*) environment vignettes or their translation. In particular, they were not Gettierized, there was no group subject, and the chances of the side effect were not manipulated. In all the analyses, we used the random-effects model with the restricted maximum likelihood (REML) heterogeneity estimator (Viechtbauer 2010), which is available in R metafor package. For each study (pair of negative and non-negative conditions) we calculated Cohen's *d* (Standardized Mean Difference, see Appendix A for details of the conversion procedure).

Publication bias. It is well known that research in which some effect is observed, especially a sizable one, is more likely to be published than research with no or slight differences between conditions (see Dickersin 2005). If ESEE were not observed in many unpublished studies then our data set would be biased. Consequently, the results of the meta-analysis would not tell us much about the true magnitude of the effect.[2]

Three methods were used to assess a possible bias in the meta-analysis: visual inspection of funnel plots, formal statistical test for the asymmetry of funnel plots (Egger *et al.* 1997; Sterne and Egger 2005) and p-curve analysis (Simonsohn *et al.* 2014*a*, *b*). Let us briefly explain the first method. In a funnel plot, effect sizes of the studies are plotted on the *x*-axis and the standard errors of these effect sizes (precision of the studies) are plotted on the *y*-axis. If there is no publication bias, studies will be distributed symmetrically around the meta-analytic estimate of the effect size: more precise studies will be clustered more tightly while less detailed studies more loosely. Publication bias can be observed as an asymmetry in the distribution of effect sizes. It will be especially pronounced towards the bottom of the plot where small (less precise) studies are represented.

In cases where the asymmetry was present (i.e. it was confirmed by the formal test or by visual inspection of the funnel plot), we estimated its impact on the meta-analytic effect size by means of two further methods: the outlier-removal and the Trim-and-Fill method. The first method of evaluating the severity of the potential bias was to exclude outlier studies from further analysis. A study was taken to be an outlier if the confidence interval of the effect size in that study did not overlap with the confidence interval of the meta-analytic estimate obtained from all studies. In the Trim-and-Fill approach (Duval and Tweedie 2000a, *b*), one first removes those studies that cause the funnel plot asymmetry. The trimmed funnel plot is then used to calculate the estimated true effect. The removed studies are then brought back and mirrored onto the other side of the funnel plot. The meta-analytic effect size and the confidence interval with the imputed "fictional" studies is then computed.

Finally, we used *p*-curve analysis, which is designed to detect *p*-hacking and publication bias. It assumes that if studies have evidential value (i.e. a true effect exists),

[2]Language is another factor that contributes to bias. Studies in English are much easier to find and are published more often.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

p-values smaller than 0.05 should have a right-skewed distribution. We used the *dmetar* R package (Harrer *et al.* 2019), which offers two statistical tests: the right-skewness test (which determines whether evidential value is present) and the flatness test (which determines whether evidential value is absent).

Meta-regression. The second type of analysis conducted was meta-regression with selected study-level covariates (see Table 2) as categorical moderators (section 3.5). Following the procedure described by Assink and Wibbelink (2016), *metafor* package was used to fit the three-level meta-analytic model.

Heterogeneity. The I^2 and T statistics were used as a measure of heterogeneity. I^2 describes the proportion of variation across studies that is attributable to heterogeneity rather than chance. T is an estimate of the standard deviation of true effects. These two measures taken together allow to assess the overall variability of ESEE.

3. Results

In section 3.1, the results of our research selection and coding are summarized. We also present the results of meta-analyses conducted on all (including unpublished) studies (section 3.2), on published studies (section 3.3), on close replications of the original experiment (section 3.4), on studies that tested the third-party ESEE (section 3.5). Finally, the results of meta-regression with selected study-level covariates are shown (section 3.6).

3.1. Studies and data

We found 14 published and 5 unpublished papers that report new experimental results on ESEE (see section 2.1). Altogether, 98 studies (pairs of a negative and a positive/neutral condition, see section 2.5) were identified as eligible for analysis (see section 2.2), of which 78 were published and 19 were unpublished. Several studies reported in eligible papers had to be excluded because not all three criteria were met (the exclusions are explained in Appendix C). Altogether, ESEE was tested on 12,568 respondents.

Table 2 provides a summary of all study-level factors alongside with the coding scheme that was used in the meta-analysis. It also includes the overall number of responses[3] for each level of the variable across all studies. Appendix B provides a list of all studies included in the meta-analysis together with the information on effect sizes,[4] study-level covariates and sample sizes.

3.2. Meta-analysis of all studies

The results of the meta-analysis of all the studies are presented as a forest plot (Figure 1). The meta-analytic estimate of the effect size ($d = 0.64$, 95% CI: 0.55–0.72) is slightly

[3]In Table 2, we report the number of responses, not the number of participants. For two experiments (Dalbauer and Hergovich 2013; Ziółkowski *et al.* Ms *a*, *b*), the number of participants is lower than the number of responses. In the former case, the same subjects were given the vignettes of two different studies. In the latter case, the subjects were asked two knowledge questions, which we treated as two studies.

[4]Guided by a reviewer's suggestion we included both Cohen's *d* and another measure of the effect size, viz. the difference between the two conditions expressed in percentage of the total scale (e.g. the difference of 2.0 on the 1–7 scale amounts to 33.3%). For studies that used dichotomous scale, the reported effect size is the difference in proportions of positive answers between two conditions (e.g. in the case of 60% of knowledge attributions in the "negative" condition and 40% in the "positive" condition, the reported effect size is 20%).

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Table 2. Study-level factors and coding scheme.

Factor	Values (total number of responses) – descriptive value (papers included)
1a. Salient	**0 (n = 10151) "true"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Dalbauer and Hergovich 2013; Beebe 2013; Beebe and Shea 2013; Turri 2014, Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Paprzycka-Hausman *et al*. Ms; Zaręba Ms; Ziółkowski *et al*. Ms *a*, *b*). **1 (n = 3676) "false"** (Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovich 2013; Buckwalter 2014; Beebe 2016; Paprzycka-Hausman 2020, 2021*c*; Wilkenfeld and Lombrozo 2020; Ryszkowska *et al*. Ms).
1b. Present	**0 (n = 11607) "true"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovich 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*, *c*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Paprzycka-Hausman *et al*. Ms; Zaręba Ms; Ziółkowski *et al*. Ms *a*, *b*). **1 (n = 2220) "false"** (Beebe and Jensen 2012; Dalbauer and Hergovich 2013; Beebe 2013; Beebe and Shea 2013; Beebe 2016; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Ryszkowska *et al*. Ms).
1c. Violated	**0 (n = 11067) "true"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovich 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*, *c*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Paprzycka-Hausman *et al.* Ms; Zaręba Ms; Ziółkowski *et al*. Ms *a*, *b*). **1 (n = 2760) "false"** (Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovich 2013; Beebe 2016; Wilkenfeld and Lombrozo 2020; Ryszkowska *et al*. Ms).
2. Scale	**0 (n = 2390) – "dichotomous"** (Beebe and Jensen 2012; Paprzycka-Hausman 2020, 2021*a*; Paprzycka-Hausman *et al*. Ms; Zaręba Ms). **1 (n =** 11437) **– "non-dichotomous"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovich 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Paprzycka-Hausman *et al*. Ms; Ryszkowska *et al*. Ms; Zaręba Ms; Ziółkowski *et al*. Ms *a*, *b*).
3. Question phrasing	**0 (n = 3591) - "knowledge probe"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Turri 2014; Paprzycka-Hausman 2020, 2021*a*; Yuan and Kim 2021; Paprzycka-Hausman *et al*. Ms; Zaręba Ms). **1 (n = 10236) – "agreement probe"** (Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovich 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*, *c*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Ryszkowska *et al*. Ms; Zaręba Ms; Paprzycka-Hausman *et al*. Ms; Ziółkowski *et al*. Ms *a*, *b*).
4. Language	**0 (n = 9722) – "English"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Wilkenfeld and Lombrozo 2020; Paprzycka-Hausman 2000, 2021*a*, *b*, *c*; Paprzycka-Hausman *et al*. Ms; Ziółkowski *et al*. Ms *a*, *b*). **1 (n = 4105) – "other"** (Dalbauer and Hergovich 2013; Yuan and Kim 2021; Ryszkowska *et al.* Ms; Zaręba Ms).
5. Chances	**0 (n = 11961) "normal chances"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovitch 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*, *c*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Paprzycka-Hausman *et al*. Ms; Ryszkowska *et al*. Ms; Zaręba Ms; Ziółkowski *et al*. Ms *a*, *b*).

(*Continued*)

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Table 2. (*Continued.*)

Factor	Values (total number of responses) – descriptive value (papers included)
	1 (n = 1866) "low chances" (Paprzycka-Hausman 2020, 2021*a*; Paprzycka-Hausman *et al.* Ms);
6. Gettierization	**0 (n = 10467) – "not Gettierized"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Dalbauer and Hergovitch 2013; Turri 2014; Beebe 2016; Wilkenfeld and Lombrozo 2020; Paprzycka-Hausman 2020, 2021*b*; Yuan and Kim 2021; Paprzycka-Hausman *et al.* Ms; Ryszkowska *et al.* Ms; Zaręba Ms; Ziółkowski *et al.* Ms *a*, *b*). **1 (n = 3360) – "Gettierized"** (Beebe and Shea 2013; Buckwalter 2014; Turri 2014; Paprzycka-Hausma 2021*c*; Yuan and Kim 2021).
7. Effect type	**0 (n = 12089) – "side effect"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovitch 2013; Buckwalter 2014; Turri 2014; Beebe 2016; Wilkenfeld and Lombrozo 2020; Paprzycka-Hausman 2021*a*, *b*, *c*; Yuan and Kim 2021; Paprzycka-Hausman *et al.* Ms; Zaręba Ms, Ziółkowski *et al.* Ms *a*, *b*). **1 (n = 1738) – "other"** (Beebe and Shea 2013; Paprzycka-Hausman 2020, 2021*a*).
8. Subject	**0 (n = 11964) "first person"** (Beebe and Buckwalter 2010; Beebe and Jensen 2012; Beebe 2013; Beebe and Shea 2013; Dalbauer and Hergovitch 2013; Buckwalter 2014; Turri 2004; Beebe 2016; Paprzycka-Hausman 2020, 2021*a*, *b*, *c*; Wilkenfeld and Lombrozo 2020; Yuan and Kim 2021; Paprzycka-Hausman *et al.* Ms; Zaręba Ms; Ziółkowski *et al.* Ms *a*, *b*). **1 (n = 1863) "third person"** (Buckwalter 2014; Beebe 2016; Yuan and Kim 2021).

smaller than the original effect size in Beebe and Buckwalter's (2010) experiment ($d = 0.74$). According to the rules of interpreting Cohen's d (Cohen 1988; Sawilowsky 2009; Gignac and Szodorai 2016; Lovakov and Agadullina 2021), the effect size is "moderate to large".

There is a high variance of effect sizes ($T = 0.34$) and a large heterogeneity of the effect estimates ($I^2 = 77.40\%$). Figure 2 shows a funnel plot for all studies included in the analysis. The distribution of medium to large studies (high-precision studies) around the summary effect size seems to be symmetrical. However, visual inspection of the plot shows a striking asymmetry for small studies (small-precision studies). The regression test for asymmetry (Egger *et al.* 1997; Sterne and Egger 2005) shows that the asymmetry of the distribution is statistically significant ($t(96) = 3.90$, $p < 0.001$).

Using the Trim-and-Fill method for bias correction (Duval and Tweedie 2000a, *b*), 17 potentially missing studies on the left side of the plot were identified. A slightly smaller summary effect size $d = 0.52$ (95% CI: 0.43–0.61) with larger heterogeneity ($I^2 = 84.19\%$, $T = 0.44$) and no detectable asymmetry in study distribution ($p = 0.68$) was obtained. The removal of outliers (22 studies) produces comparable results ($d = 0.63$, 95% CI: 0.57–0.68, $I^2 = 33.08\%$, $T = 0.13$, test for asymmetry: $p = 0.068$).[5] Figure 3 shows the results obtained by these two methods.

P-curve analysis for all studies shows no evidence of *p*-hacking or publication bias (right-skewness test: $z = -22.96$, $p < 0.001$; flatness test: $z = 15.77$, $p > 0.999$; more detailed results and plots are available in Appendix D).

[5]Although the I^2 and T statistics are substantially reduced, the outlier-removal method produces unreliable estimates of heterogeneity.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

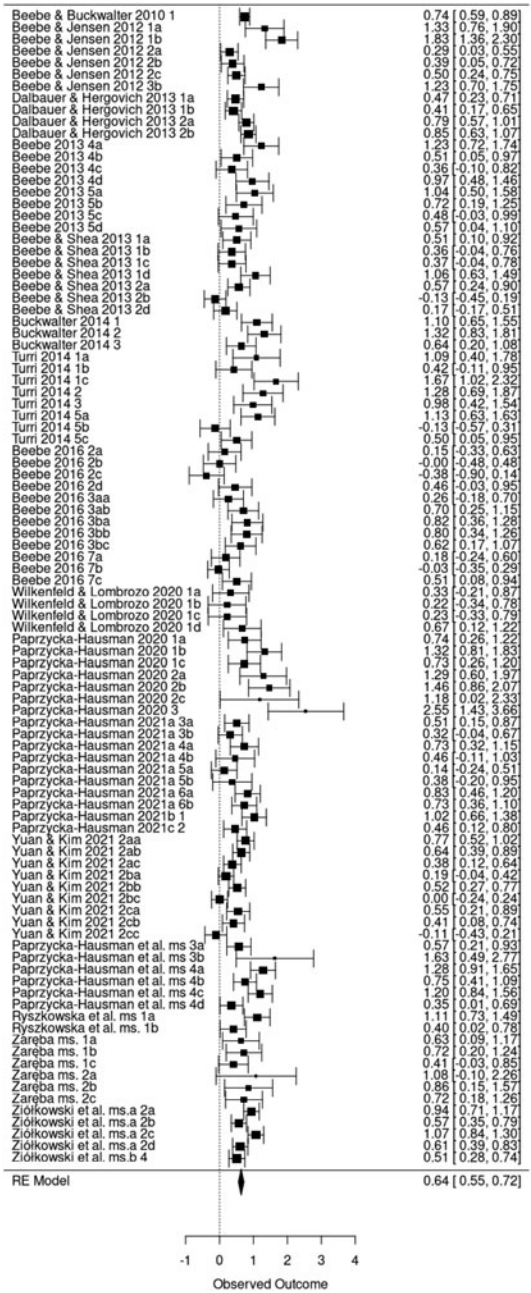


Figure 1. Forest plot for all studies on ESEE (distribution of effect sizes and confidence intervals).

3.3. Meta-analysis of all published studies

The results of the meta-analysis for all studies that were published are presented in Figure 4. The estimates are almost indistinguishable from those obtained for all studies ($d = 0.61$, 95% CI: 0.52–0.70, $I^2 = 78.42\%$, $T = 0.36$).

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

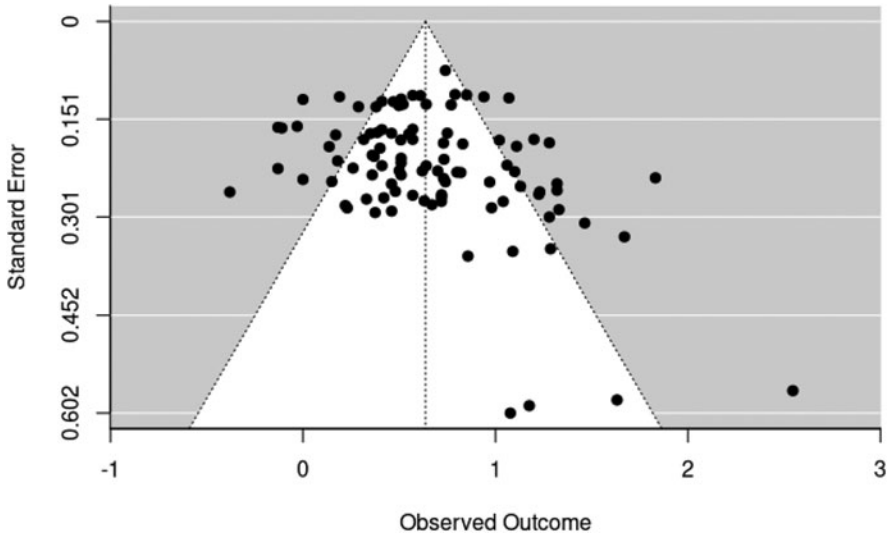


Figure 2. Funnel plot for all studies on ESEE (*x*-axis: observed effect size; *y*-axis: precision of study measured by standard error).

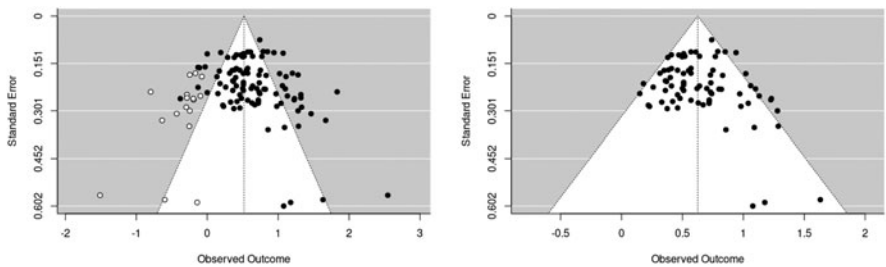


Figure 3. Bias-corrected funnel plots for all studies on ESEE (*x*-axis: observed effect size; *y*-axis: precision of study measured by standard error): Trim-and-Fill method (on the left), outlier-removal method (on the right)

The funnel plot (Figure 5) shows a statistically significant asymmetry ($t(77) = 4.25$, $p < 0.001$). Both methods for correcting asymmetry produced comparable results (Trim-and-Fill: 12 studies missing, $d = 0.50$ (95% CI: 0.39–0.60), $I^2 = 84.88\%$, $T = 0.47$; outlier removal: 17 studies; $d = 0.59$ (95% CI: 0.52 –0.65), $I^2 = 41.35\%$, $T = 0.16$).

P-curve analysis for published studies shows no evidence of *p*-hacking or publication bias (right-skewness test: $z = -19.24$, p <0.001; flatness test: $z = 12.95$, p >0.999; more detailed results and plots are available in Appendix D).

3.4. Meta-analysis of close replications

As mentioned earlier, we decided to analyse all studies that were close replications of Beebe and Buckwalter's (2010) experiment separately. We took close replications to be those experiments that used Knobe's (2003*a*) environment vignettes or their direct translations. In close replications, there was no Gettierization and the chances of the side effect were not manipulated. There were 16 such studies. The results of the meta-analysis are presented in Figure 6. The summary effect size for close replications

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

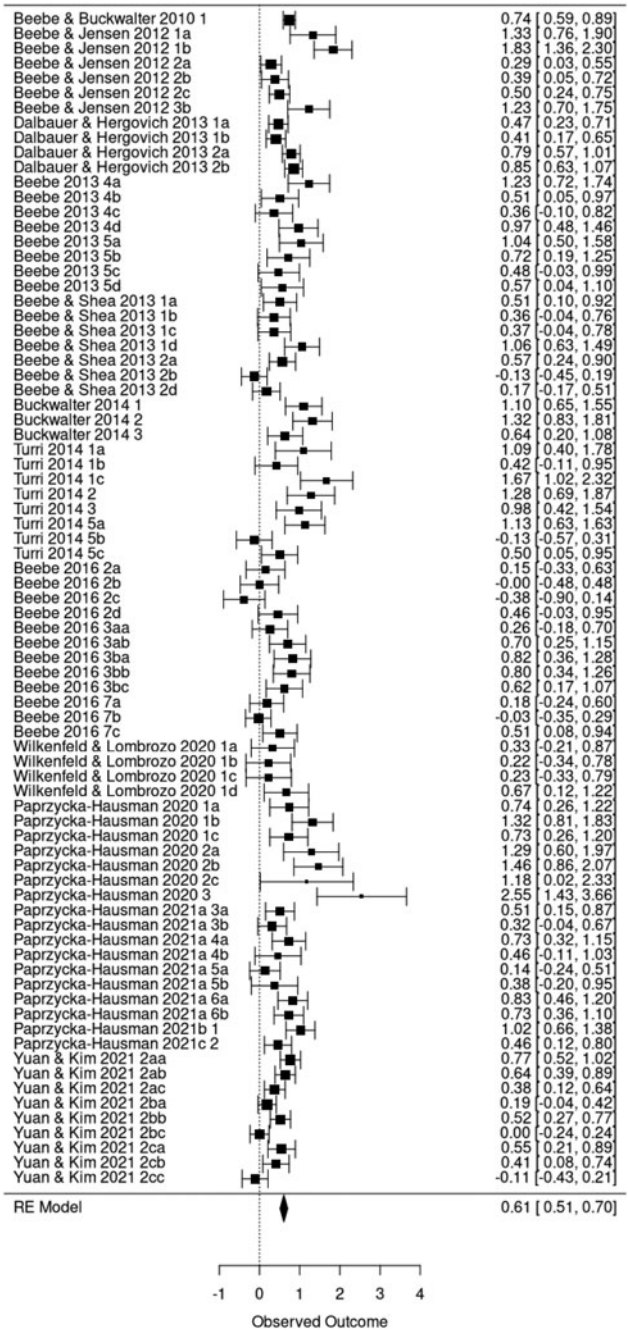


Figure 4. Forest plot for published studies on ESEE (distribution of effect sizes and confidence intervals).

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

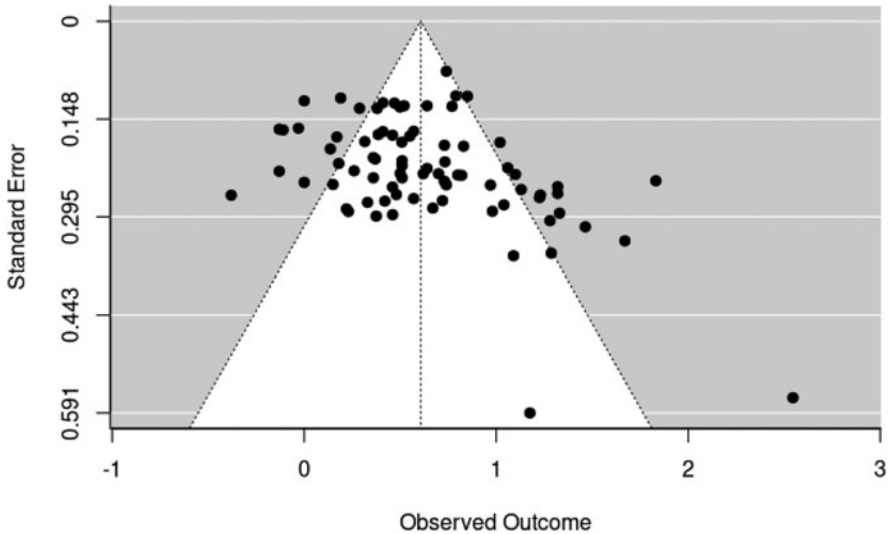


Figure 5. Funnel plot for published studies on ESEE (*x*-axis: observed effect size; *y*-axis: precision of study measured by standard error).

($d = 0.85$, 95% CI: 0.62–1.08) is considerably larger than the summary effect sizes for all studies and for all published studies.

There is a relatively high variance ($T = 0.34$) and a large heterogeneity ($I^2 = 74.27\%$) among close replications even though they are relatively homogenous in experimental design and materials used. Visual examination of the funnel plot suggests that there might be a slight positive bias of the meta-analytic effect size (see Figure 7). However, the asymmetry is not statistically significant ($p = 0.095$).

The Trim-and-Fill method (see Figure 8) produced a considerably smaller meta-analytic estimate of the effect size (5 studies potentially missing, $d = 0.66$, 95% CI: 0.42–0.91) with heterogeneity and variance considerably larger than without it ($I^2 = 83.66\%$, $T = 0.49$).

Again, *p*-curve analysis for close replication studies shows no evidence of *p*-hacking or publication bias (right-skewness test: $z = -9.96$, $p < 0.001$; flatness test: $z = 7.14$, $p > 0.999$; more detailed results and plots are available in Appendix D).

3.5. Meta-analysis of third-party ESEE

Third-party ESEE was tested in 16 studies (cf. section 2.5). The overall meta-analytic estimate of the effect size was still statistically significant but it was considerably smaller than in our previous analysis ($d = 0.31$, 95% CI: 0.12–0.50, $p = 0.003$). The results of the analysis are presented in Figure 9. There is no evidence for publication bias in these studies ($p = 0.38$). P-curve analysis for third-party ESEE studies shows no evidence of p-hacking or publication bias (right-skewness test: $z = -4.48$, $p < 0.001$; flatness test: $z = 2.21$, $p = 0.987$; more detailed results and plots are available in Appendix D).

3.6. Meta-regression with study-level covariates

The results for meta-regression are presented in Table 3. We have decided to run separate meta-regressions for all three norm coding schemes: *Present*, *Violated* and *Salient*

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

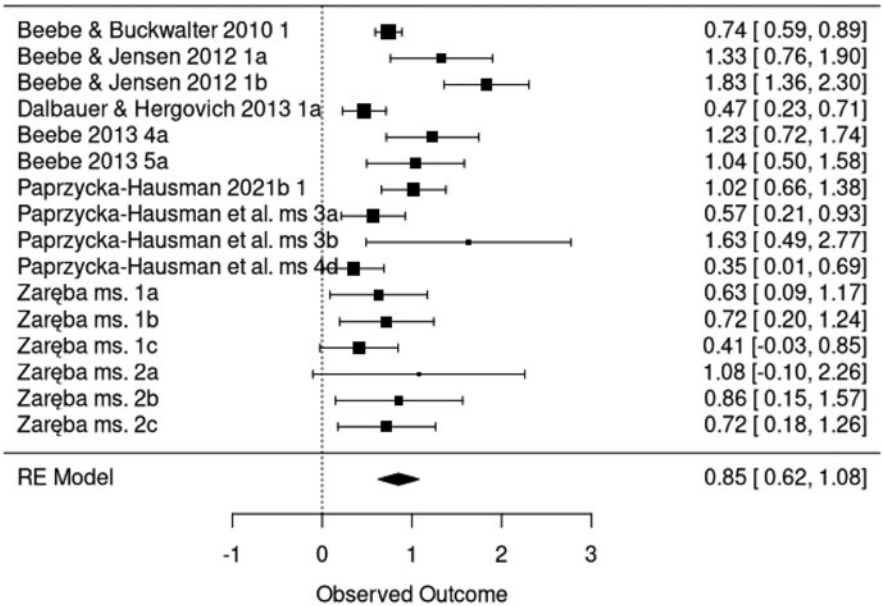


Figure 6. Forest plot for close replications of ESEE (distribution of effect sizes and confidence intervals).

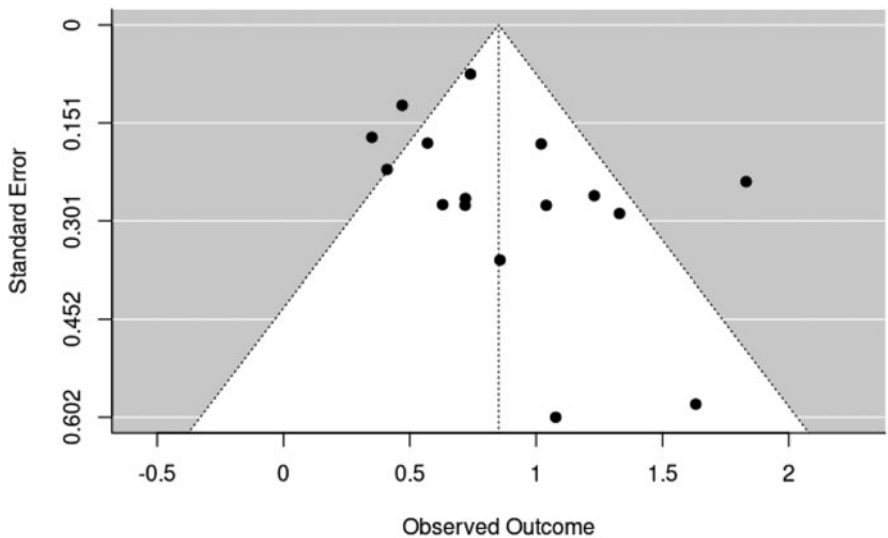


Figure 7. Funnel plot for close replications of ESEE (*x*-axis: observed effect size; *y*-axis: precision of study measured by standard error).

(see section 1.1). All three models are included in Table 3. In all models, the same two factors affect the size of the asymmetry in knowledge attributions. They both decrease the asymmetry. The effect size is decreased when (1) a non-moral norm is violated, and (2) when participants assess knowledge of a third party rather than the agent.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

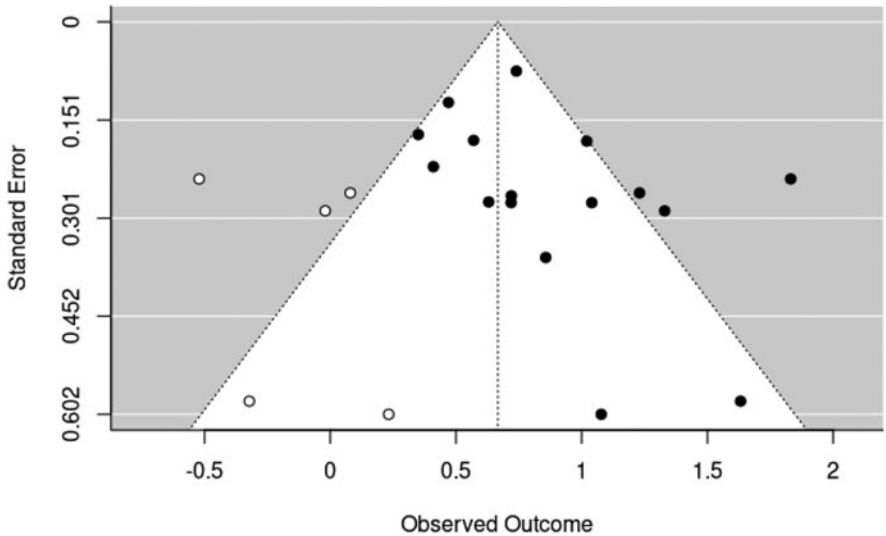


Figure 8. Bias-corrected (Trim-and-Fill) funnel plots for close replications of ESEE (*x*-axis: observed effect size; *y*-axis: precision of study measured by standard error).

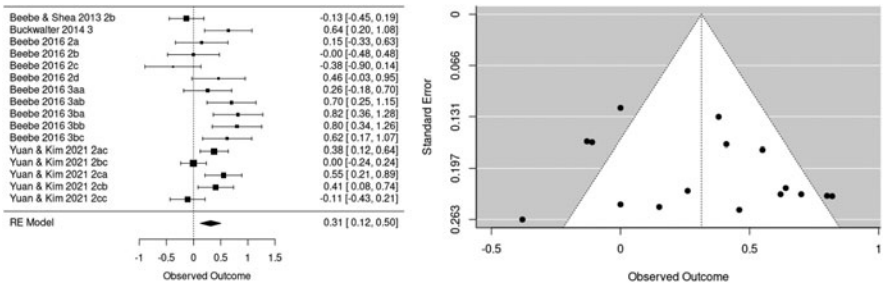


Figure 9. On the left: forest plot for third-party ESEE studies (distribution of effect sizes and confidence intervals); on the right: funnel plot for third-party ESEE studies (*x*-axis: observed effect size; *y*-axis: precision of study measured by standard error).

The model that uses the *Present* coding scheme also indicates that Gettierization has a marginally significant negative impact on the effect size. The test for all moderators is statistically significant for all three models (see "Test for all moderators" in Table 3). It should be pointed out that much of the variance remains unexplained (see "Test for residual heterogeneity" in Table 2). We compared the goodness of fit of meta-regression models using the Akaike information criterion (AIC). The differences were relatively small but the *Present* coding scheme turned out to produce the best-fitted model.

4. Discussion

Our analysis established that ESEE is a rather robust phenomenon with conventionally interpreted moderate to large effect size. It should be stressed that the rules of thumb for interpreting effect sizes (Cohen 1988; Sawilowsky 2009; Gignac and Szodorai 2016;

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Table 3. Three meta-regression models (each uses a different norm coding scheme: *Salient*, *Violated*, *Present*).

	Salient		*Violated*		*Present*	
Predictor	Cohen's *d*	*p* value	Cohen's *d*	*p* value	Cohen's *d*	*p* value
(intercept)	0.775	<0.0001	0.782	<0.0001	0.783	<0.0001
Non-moral norm	−0.264	0.007	−0.246	0.018	−0.313	0.004
Third-party subject	−0.347	0.002	−0.334	0.003	−0.341	0.002
Gettierized story	−0.132	0.213	−0.167	0.128	−0.180	0.096
Low chances	0.115	0.462	0.176	0.249	0.175	0.246
Other effect	0.023	0.882	−0.129	0.385	−0.120	0.413
Non-English Language	−0.075	0.456	−0.049	0.622	−0.025	0.801
Not published	0.002	0.982	0.004	0.968	−0.003	0.979
"Agreement" phrasing	0.039	0.814	0.017	0.920	0.013	0.934
Dichotomous scale	0.002	0.991	0.009	0.965	−0.002	0.994
Test for all moderators	$F(9, 88) = 3.460$ $p = 0.001$		$F(9, 88) = 3.228$ $p = 0.002$		$F(9, 88) = 3.626$ $p < 0.001$	
Test for residual heterogeneity	$QE(88) = 284.756$ $p < 0.0001$		$QE(88) = 286.456$ $p < 0.0001$		$QE(88) = 282.596$ $p < 0.0001$	
AIC	111.787		113.565		110.751	

Lovakov and Agadullina 2021) are tailored to disciplines such as social psychology. Even if the effect size could be interpreted as large, it would still be an open question whether it would be philosophically significant. Consider the Knobe effect. For Knobe's (2003*a*) study, the summary effect size exceeds that for ESEE ($d = 1.53$ when the method of converting Odds Ratio to Cohen's *d* is used). Of course, to compare ESEE with the Knobe effect, a meta-analysis of the latter is needed. Still, this difference is notable.

The analysis of funnel plots provides evidence of a slight asymmetry in the distribution of studies around the summary effect. It is arguable, however, that this does not indicate a publication bias in the research on ESEE or, more generally, in experimental philosophy. Rather, the asymmetry is produced by a few low-precision studies that obtained large effects. In addition, *p*-curve analysis does not indicate any presence of *p*-hacking or systematic publication bias. It is thus reasonable to treat those low-precision studies as outliers rather than an indication of a more general tendency. After an appropriate correction, meta-analytic estimates did not change dramatically. Indeed, ESEE reached statistical significance in the vast majority of studies, including those that are yet unpublished. ESEE can be thus considered to be a robust phenomenon: it is present in experimental setups that use different experimental materials and procedures.

In all our meta-analyses, there is a large heterogeneity of effect sizes indicated by high values of I^2 and T statistics. Perhaps such variability could be expected since our sample of studies consists of experiments that differ in many respects. However, very large dispersion estimates were obtained even for close replications of Beebe and

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Buckwalter's (2010) experiment, i.e. for studies that differ only in language, question phrasing or response scale. These findings indicate that while ESEE is robust, it is also quite variable. At this point, the sources of its variability are not fully known. One might perhaps try to explain it by factors such as the demography of the sample (e.g. university students vs mTurkers), the mode of presentation (online vs pen and paper) or the method of recruiting participants (crowdsourcing services vs volunteers).

A notable and interesting result of the analysis is the presence of third-party ESEE. So far, the data were equivocal. The effect was statistically significant in eight studies but not significant in seven studies. Thus the presence of the effect could not be established by a mere study count. Our analysis showed that despite the fact that statistical significance was not reached in several studies, the overall meta-analytic effect is statistically significant. The effect size should be interpreted as "small to moderate" (Cohen 1988; Sawilowsky 2009; Gignac and Szodorai 2016; Lovakov and Agadullina 2021). Notably, rather large sample sizes were used in several experiments that attempted to replicate Buckwalter's (2014) finding. In two of such studies, there was a non-significant difference in the opposite direction (Beebe 2016; Yuan and Kim 2021). Replication failure in this case cannot be explained in terms of the low power of the studies in question. Third-party ESEE should be investigated further though the results of our analysis confirm its presence.

Meta-regression analysis shows that two out of eight study-level covariates affect the magnitude of ESEE: the subject of the knowledge attribution (agent vs observer) and the type of the norm present/violated/salient (moral vs non-moral). The results also suggest that Gettierization might play a small role (marginally significant p value).

The change of the subject of knowledge attribution from the agent to the observer (third-party) decreases the magnitude of the asymmetry. The asymmetry tends to be smaller when knowledge is attributed to the observer (but, as noted earlier, it is still present). In the literature, there are competing explanations of the third-party effect. The issue has been made even more complicated by the poor replicability of Buckwalter's (2014) results. Buckwalter introduced the third-party condition to eliminate a possibility raised by Alicke (2008) that the folk concept of knowledge is sensitive to moral evaluation because people want to hold the protagonist responsible for the negative outcome of his actions. Buckwalter argued that Alicke's solution is unsatisfactory since the asymmetry is present even in the attribution of knowledge to a subject who is merely a witness of events rather than their agent. Our meta-analysis resolves two issues. First, in normative contexts, there is indeed an asymmetry in the attribution of knowledge to subjects who are merely witnesses of events but not their agents. Second, there is a significant difference in the attribution of knowledge to a protagonist who brings about a result and a witness who does not actively participate in producing the result. This alone, of course, does not resolve the question which account of ESEE is more feasible. It may indicate that explanations that can account for the third-party ESEE are preferrable but there is also a possibility that the assumption that ESEE should be explained by one account is false. It might be the case that different accounts can explain some part of the large heterogeneity of the effect sizes that we have observed. Nevertheless it is important to stress that third-party ESEE should not be overlooked in attempts to explain ESEE.

The meta-analysis also shows that the asymmetry is generally smaller in those scenarios where a non-moral rather than a moral norm is present, violated or made salient by the question. Clearly, these two different types of norms differ in their scope of application, in what justifies them, and in people's reaction to their violation. At first

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

glance, a violation of moral standards poses a more serious threat to human well-being. It is also more strongly condemned than a violation of conventional norms, which typically do not have such severe consequences. Moreover, moral standards seem to be perceived by people as more general (universal) in scope and more 'objective' than non-moral ones (for critical discussion, see O'Neill 2017). According to some empirical studies (Van Bavel *et al.* 2012), moral evaluations are faster, more extreme, and more strongly associated with universal prescription (such as that nobody or everybody should engage in action) than non-moral evaluations. This suggests that moral norms are more prominent than non-moral ones, which might explain the observed impact.

Our analysis had several interesting negative findings. There were no effects of scale, effect type (side effect vs other), manipulation of chances, language or question phrasing. Let us consider them in turn.

As it turned out, the results obtained using dichotomous and non-dichotomous scales are comparable. The concern about scales was first raised at the beginning of research on ESEE by Beebe and Jensen (2012). They found that ESEE was present when different Likert scales were used as well as when a dichotomous scale was used. Our results confirm and strengthen their early finding.

We found no general difference between studies that used side-effect stories and those that used other stories. Recall that in Butler-type scenarios as well as in some Gettierized studies, the stories are about the main effect rather than about a side effect. This finding provides further evidence for the claim defended earlier (section 1.7) that the same kind of phenomenon is manifested in both types of scenarios. It might also be taken to support the suggestion that the name "side-effect effect" has been coined prematurely.

That no effect of manipulation of chances was found deserves some reflection. When the event is said to occur with a high probability (or when the probability is not given), people are more inclined to attribute knowledge that the event will occur in norm-violation scenarios than in norm-conformity scenarios. Prima facie, when the event is said to occur with a low probability, people ought not to be inclined to attribute knowledge that it will occur in either condition. In other words, it would be rational to expect that the manipulation of chances will decrease the effect. The fact that the manipulation of chances does not reduce the effect thus calls for an explanation. The finding seems to undermine probability-based accounts of ESEE (e.g. Dalbauer and Hergovich 2013).

Language does not have a statistically significant effect either. It would be premature, however, to generalize from our results and claim that ESEE is cross-linguistically or cross-culturally universal. Our sample of studies included research in only five languages – English, German (both are Germanic languages), Polish, Korean and Chinese. To establish universality for ESEE, one would need to conduct a more systematic replication of the original study in different languages and cultures. For now, however, ESEE seems to be cross-culturally robust. It thus makes its way onto a growing list of culturally universal epistemic phenomena such as Gettier intuitions (Machery *et al.* 2017, 2018) or the lack of impact of stakes on knowledge attribution (Rose *et al.* 2019).

Finally, the phrasing of the question had no impact on the effect size. From a philosophical point of view, the difference between an agreement probe ("To what extent do you agree …" with different levels of agreement to be marked on the scale) and a knowledge probe ("Did *S* know that …" with different levels of knowledge to be marked on the scale) is obvious. These two formulations operationalize different constructs. If knowledge were gradable, it would be possible to attribute a "weak" level of knowledge

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

with certainty, just as it would be possible to attribute a "high" level of knowledge with hesitation (cf. Zyglewicz and Maćkiewicz 2019 for discussion of two types of gradability). The results of the meta-analysis indicate that as far as experimental studies are concerned, this difference does not produce different patterns of responses.

5. Conclusions

Our analysis shows that ESEE is a robust though variable effect with a moderate to large effect size. Our meta-analyses show a large heterogeneity of effect sizes for all studies considered but, remarkably, also for close replications of Beebe and Buckwalter's (2010) experiment. In general, there is no evidence of publication bias in the research on ESEE.

A notable and interesting result of the analysis is the presence of third-party ESEE. Despite the fact that statistical significance was not reached in several studies, the overall meta-analytic effect is statistically significant. The effect size of third-party ESEE should be interpreted as "small to moderate".

Meta-regression analysis shows that two out of eight study-level covariates affect the magnitude of ESEE: the subject of the knowledge attribution (agent vs observer) and the type of the norm present/violated/salient (moral vs non-moral). The results also suggest that Gettierization might play a small role (marginally significant p value).

Our analysis also shows that several factors did not have a statistically significant influence on the effect size. Some of these factors pertained to the language or the method of study: the type of scale used (dichotomous vs non-dichotomous) and question phrasing (agreement vs knowledge probe). The other factors that turned out not to affect the effect size pertained to the content of the stories. Whether the story was about a side effect or not turns out not to influence the effect size, which suggests that the name for ESEE has been coined prematurely. Interestingly, the manipulation of chances (high/normal vs low) also does not have an impact on the effect size. The size of the asymmetry in the attributions of knowledge is present even if the chances of the event are described as low.[6]

References

Alfano M., Beebe J.R. and Robinson B. (2012). 'The Centrality of Belief and Reflection in Knobe-effect Cases: A Unified Account of the Data.' *The Monist* **95**(2), 264–89.

Alicke M. (2008). 'Blaming Badly.' *Journal of Cognition and Culture*, **8**(1–2), 179–86.

Assink M. and Wibbelink C.J. (2016). 'Fitting Three-level Meta-analytic Models in R: A Step-by-step Tutorial.' *Quantitative Methods for Psychology* **12**(3), 154–74.

Van Bavel J.J., Packer D.J., Haas I.J. and Cunningham W.A. (2012). 'The Importance of Moral Construal: Moral versus Non-Moral Construal Elicits Faster, More Extreme, Universal Evaluations of the Same Actions.' *PLoS ONE* **7**(11), e48693.

Beebe J.R. (2013). 'A Knobe Effect for Belief Ascriptions.' *Review of Philosophy and Psychology* **4**(2), 235–58.

Beebe J.R. (2016). 'Do Bad People Know More? Interactions between Attributions of Knowledge and Blame.' *Synthese* **193**(8), 2633–57.

Beebe J.R. and Buckwalter W. (2010). 'The Epistemic Side-Effect Effect.' *Mind & Language* **25**(4), 474–98.

[6]We thank our students Weronika Bakun, Marcin Brejnak, Aleksandra Szwech, Aleksandra Wojsz and Michał Węgierski for assistance in study selection and data extraction. The project has been financed by grant (2018/29/B/HS1/02861) from the National Science Centre, Poland.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Beebe J.R. and Jensen M. (2012). 'Surprising Connections between Knowledge and Action: The Robustness of the Epistemic Side-effect Effect.' *Philosophical Psychology* **25**(5), 689–715.

Beebe J.R. and Shea J. (2013). 'Gettierized Knobe Effects.' *Episteme* **10**(3), 219–40.

Borenstein M., Hedges L.V., Higgins J.P. and Rothstein H.R. (2009). *Introduction to Meta-analysis*. Chichester: Wiley.

Buckwalter W. (2014). 'Gettier Made ESEE.' *Philosophical Psychology* **27**(3), 368–83.

Butler R.J. (1978). 'Report on Analysis "Problem" No. 16.' *Analysis* **38**(3), 113–14.

Charness G., Gneezy U. and Kuhn M.A. (2012). 'Experimental Methods: Between-subject and Within-subject Design.' *Journal of Economic Behavior & Organization* **81**(1), 1–8.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. London: Routledge.

Dalbauer N. and Hergovich A. (2013). 'Is What is Worse More Likely? – The Probabilistic Explanation of the Epistemic Side-Effect Effect.' *Review of Philosophy and Psychology* **4**(4), 639–57.

Dickersin K. (2005). 'Publication Bias: Recognizing the Problem, Understanding its Origins and Scope, and Preventing Harm.' In H.R. Rothstein, A.J. Sutton, M. Borenstein (eds), *Publication Bias in Meta-analysis: Prevention, Assessment and Adjustments*, pp. 11–33. Chichester: Wiley.

Duval S. and Tweedie R. (2000*a*). 'Trim-and-Fill: A Simple Funnel-plot–based Method of Testing and Adjusting for Publication Bias in Meta-analysis.' *Biometrics* **56**(2), 455–63.

Duval S. and Tweedie R. (2000*b*). 'A Nonparametric "Trim-and-Fill" Method of Accounting for Publication Bias in Meta-analysis.' *Journal of the American Statistical Association* **95**(449), 89–98.

Egger M., Smith G.D., Schneider M. and Minder C. (1997). 'Bias in Meta-analysis Detected by a Simple, Graphical Test.' *British Medical Journal* **315**(7109), 629–34.

Gignac G.E. and Szodorai E.T. (2016). 'Effect Size Guidelines for Individual Differences Researchers.' *Personality and Individual Differences* **102**, 74–8.

Harman G. (1976). 'Practical Reasoning.' *Review of Metaphysics* **29**(3), 431–63.

Harrer M., Cuijpers P., Furukawa T. and Ebert D.D. (2019). *dmetar: Companion R Package for the Guide 'Doing Meta-Analysis in R'*. R package version 0.0.9000. http://dmetar.protectlab.org/.

Hindriks F. (2008). 'Intentional Action and the Praise-Blame Asymmetry.' *Philosophical Quarterly* **58** (233), 630–41.

Hindriks F. (2011). 'Control, Intentional Action, and Moral Responsibility.' *Philosophical Psychology* **24**(6), 787–801.

Holton R. (2010). 'Norms and the Knobe Effect.' *Analysis* **70**(3), 417–24.

Knobe J. (2003*a*). 'Intentional Action and Side Effects in Ordinary Language.' *Analysis* **63**(3), 190–4.

Knobe J. (2003*b*). 'Intentional Action in Folk Psychology. An Experimental Investigation.' *Philosophical Psychology* **16**(2), 309–24.

Knobe J. (2004). 'Folk Psychology and Folk Morality: Response to Critics.' *Journal of Theoretical and Philosophical Psychology* **24**(2), 270–9.

Knobe J. (2007). 'Reason Explanation in Folk Psychology.' *Midwest Studies in Philosophy* **31**, 90–106.

Knobe J. (2010). 'Person as Scientist, Person as Moralist.' *Behavioral and Brain Sciences* **33**(4), 315–29.

Knobe J. and Mendlow G.S. (2004). 'The Good, the Bad and the Blameworthy: Understanding the Role of Evaluative Reasoning in Folk Psychology.' *Journal of Theoretical and Philosophical Psychology* **24**(2), 252–8.

Knobe J., Buckwalter W., Nichols S., Robbins P., Sarkissian H. and Sommers T. (2012). 'Experimental Philosophy.' *Annual Review of Psychology* **63**(1), 81–99.

Lovakov A. and Agadullina E.R. (2021). 'Empirically Derived Guidelines for Effect Size Interpretation in Social Psychology.' *European Journal of Social Psychology* **51**(3), 485–504.

Machery E., Stich S., Rose D., Chatterjee A., Karasawa K., Struchiner N., Sirker S., Usui N. and Hashimoto T. (2017). 'Gettier across Cultures.' *Noûs* **51**(3), 645–64.

Machery E., Stich S., Rose D., Chatterjee A., Karasawa K., Struchiner N., Sirker S., Usui N. and Hashimoto T. (2018). 'Gettier was Framed.' In M. Mizumoto, S.P. Stich and E.S. McCready (eds), *Epistemology for the Rest of the World*, pp. 123–48. Oxford: Oxford University Press.

Nadelhoffer T. (2004). 'The Butler Problem Revisited.' *Analysis* **64**(3), 277–84.

Nichols S. and Ulatowski J. (2007). 'Intuitions and Individual Differences. The Knobe Effect Revisited.' *Mind and Language* **22**(4), 346–65.

O'Neill E. (2017). 'Kinds of Norms.' *Philosophy Compass* **12**(5), 12:e12416.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Paprzycka-Hausman K. (2020). 'Knowledge of Consequences: An Explanation of the Epistemic Side-Effect Effect.' *Synthese* **197**(12), 5457–90.

Paprzycka-Hausman K. (2021*a*). 'Przypisania wiedzy w kontekstach niskiego prawdopodobieństwa. Studia eksperymentalne.' In K. Paprzycka-Hausman, K. Kuś and B. Maćkiewicz (eds), *Wyjaśnienia efektu Knobe'a i problemu Butlera. Studium teoretyczne i eksperymentalne efektu Knobe'a i problemu Butlera. Vol. 1*, pp. 108–72. Semper.

Paprzycka-Hausman K. (2021*b*). 'Hipoteza zaniechań a model heurystyk doksastycznych.' In K. Paprzycka-Hausman, K. Kuś and B. Maćkiewicz (eds), *Wyjaśnienia efektu Knobe'a i problemu Butlera. Studium teoretyczne i eksperymentalne efektu Knobe'a i problemu Butlera. Vol. 1*, pp. 68–80. Semper.

Paprzycka-Hausman K. (2021*c*). 'Efekt Epistemiczny a problem Gettiera w świetle hipotezy zaniechaniowej.' In K. Paprzycka-Hausman, K. Kuś and B. Maćkiewicz (eds), *Wyjaśnienia efektu Knobe'a i problemu Butlera. Studium teoretyczne i eksperymentalne efektu Knobe'a i problemu Butlera. Vol. 1*, pp. 173–200. Semper.

Paprzycka-Hausman K., Kuś K., Maćkiewicz B. and Zaręba M. (Ms) *Replication of ESSE During a Global Pandemic*. University of Warsaw.

Pinillos N.Á., Smith N., Nair G.S., Marchetto P. and Mun C. (2011). 'Philosophy's New Challenge: Experiments and Intentional Action.' *Mind and Language* **26**(1), 115–39.

Robinson B., Stey P. and Alfano M. (2015). 'Reversing the Side-effect Effect: The Power of Salient Norms.' *Philosophical Studies* **172**(1), 177–206.

Rose D., Machery E., Stich S., Alai M., Angelucci A., Berniūnas R. … and Zhu J. (2019). 'Nothing at Stake in Knowledge.' *Noûs* **53**(1), 224–47.

Ryszkowska M., Skrzecz K. and Wiśniewska U. (Ms). *The Group Knobe Effect – Data from Polish Speakers*. University of Warsaw.

Sawilowsky S.S. (2009). 'New Effect Size Rules of Thumb.' *Journal of Modern Applied Statistical Methods* **8** (2), 26.

Simonsohn U., Nelson L.D. and Simmons J.P. (2014*a*). 'P-curve: a Key to the File-drawer.' *Journal of Experimental Psychology* **143**(2), 534.

Simonsohn U., Nelson L.D. and Simmons J.P. (2014*b*). 'P-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results.' *Perspectives on Psychological Science* **9**(6), 666–81.

Sterne J.A.C. and Egger M. (2005). 'Regression Methods to Detect Publication and Other Bias in Meta-analysis.' In H.R. Rothstein, A.J. Sutton and M. Borenstein (eds), *Publication Bias in Meta-analysis: Prevention, Assessment, and Adjustments*, pp. 99–110. Chichester: Wiley.

Turri J. (2014). 'The Problem of ESEE Knowledge.' *Ergo* **1**(4), 101–27.

Viechtbauer W. (2010). 'Conducting Meta-analyses in R with the metafor Package.' *Journal of Statistical Software* **36**(3), 1–48.

Wilkenfeld D.A. and Lombrozo T. (2020). 'Explanation Classification Depends on Understanding: Extending the Epistemic Side-Effect Effect.' *Synthese* **197**(6), 2565–92.

Yuan Y. and Kim M. (2021). 'Cross-Cultural Convergence of Knowledge Attribution in East Asia and the US.' *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-021-00523-y.

Zaręba M. (Ms). 'Epistemic Side Effect Effect in Polish Language.' University of Warsaw.

Ziółkowski A.A. (2017). 'Experimenting on Contextualism: Between-Subjects vs. Within-Subjects.' *Teorema: Revista Internacional de Filosofía*, 139–162.

Ziółkowski A., Tałasiewicz M., Tarnowski M. and Jusińska Z. (Ms *a*). 'Replication and Extension of the Michael and Szigeti's (2019) Experiment 4.'

Ziółkowski A., Tałasiewicz M., Tarnowski M. and Jusińska Z. (Ms *b*). 'Replication and Extension of the Michael and Szigeti's (2019) Experiment 2b.'

Zyglewicz T. and Maćkiewicz B. (2019). 'Objective and Epistemic Gradability: Is the New Angle on the Knobe Effect Empirically Grounded?' *Philosophical Psychology* **32**(2), 234–56.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Appendix A. Procedure used for data imputation and conversion

Computing Cohen's *d*

For studies that used a non-dichotomous scale:

1. If Cohen's *d* was reported, we used the magnitude reported.
2. If means, standard deviations and sample sizes were reported for each condition, we computed Cohen's *d* from the statistics reported. If standard deviations were not reported, they were computed from standard errors and sample sizes, if available.
3. If the *t* test statistic and the sample size for each condition were reported, we used the following conversion formula:

$$d = t\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

4. If the *t* test statistic and the total sample size were reported, we used the following conversion formula:

$$d = \sqrt{\frac{2\mathrm{t}}{\sqrt{\mathrm{n}}}}$$

5. If only the total sample size was reported but means and standard deviations were provided for each group, we computed Cohen's *d* from the reported statistics on the assumption that number of participants was equal for each condition ($n_1 = n_2$).

For studies that used a dichotomous scale:

6. If sample sizes as well as either counts or percentages of responses for each condition were reported, we computed the odds ratio and converted it to Cohen's *d* by means of the following conversion formula:

$$d = \ln \mathrm{OR} \frac{\sqrt{3}}{\pi}$$

7. If the percentage of responses was reported for each condition but only the total sample size was provided, we assumed that both groups are of equal size and used method (6) accordingly.

Computing the sampling variance of Cohen's *d*

For studies that used a non-dichotomous scale:

8. If separate sample sizes for each condition were reported, we computed the sampling variance of Cohen's *d* by means of the following formula:

$$\mathrm{V}_d = \frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

9. If only the total sample size was reported, we used formula from (8) assuming that the number of participants was equal for each condition ($\mathrm{n}_1 = \mathrm{n}_2$).

For studies that used a dichotomous scale:

10. We calculated approximate variance using the formula:

$$V_{LogOddsRatio} \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

and then converted it to V_d using the following formula:

$$V_d = V_{LogOddsRatio} \frac{3}{\pi^2}$$

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Appendix B. Studies included in the meta-analysis

Note: Values in the "% difference" columns refer to the difference between the two conditions expressed in percentage of the total scale (e.g. the difference of 2.0 on the 1–7 scale amounts to 33.3%). For studies that used dichotomous scale, the reported effect size is the difference in proportions of positive answers between two conditions (e.g. in case of 60% of knowledge attributions in the "negative" condition and 40% in the "positive" condition, the reported effect size is 20%).

Paper	Study	Subgroup	Violated	Present	Salient	Scale	Q Phrasing	Language	Chances	Gettierization	Effect	Subject	% difference	d	N
Beebe and Buckwalter 2010	1	(original study)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	22.3 %	0.740	749
Beebe and Jensen 2012	1	a (dich. replication)	true	true	true	dich.	true	English	normal chances	not Gettierized	side effect	first	52 %	1.330	87
		b (non-dich. replication)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	50.3 %	1.831	98
	2	a ("movies" story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	9.7 %	0.288	234
		b ("sales" story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	11.7 %	0.386	139
		c ("nazi" story)	false	true	false	non-dich	true	English	normal chances	not Gettierized	side effect	first	14.2 %	0.498	246
	3b	(gang leader)	true	true	true	dich.	true	English	normal chances	not Gettierized	side effect	first	34 %	1.225	167
Dalbauer and Hergovich 2013	1	a ("profit" story)	true	true	true	non-dich	true	other	normal chances	not Gettierized	side effect	first	9.1 %	0.470	269
		b ("guideline" story)	true	true	true	non-dich	true	other	normal chances	not Gettierized	side effect	first	8.2 %	0.410	269
	2	a ("movie" story)	false	false	false	non-dich	false	other	normal chances	not Gettierized	side effect	first	15.7 %	0.790	340
		b ("reorganization" story)	false	false	false	non-dich	false	other	normal chances	not Gettierized	side effect	first	16.5 %	0.850	340
Beebe 2013	4	a ("environment" story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	29.2 %	1.230	69
		b ("movies" story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	12.3 %	0.510	74
		c ("sales" story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	11.2 %	0.360	73
		d ("nazi" story)	false	true	false	non-dich	true	English	normal chances	not Gettierized	side effect	first	29.2 %	0.970	73
	5	a ("environment" story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	26.2 %	1.040	59
		b ("movies" story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	24.2 %	0.720	58
		c ("sales" story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	14 %	0.480	60
		d ("nazi" story)	false	true	false	non-dich	true	English	normal chances	not Gettierized	side effect	first	15.7 %	0.570	58
Beebe and Shea 2013	1	a ("water" story Gettierized)	true	true	false	non-dich	true	English	normal chances	Gettierized	side effect	first	15.3 %	0.510	93
		b ("movies" story Gettierized)	false	false	false	non-dich	false	English	normal chances	Gettierized	side effect	first	10.2 %	0.360	96
		c ("sales" story Gettierized)	false	false	false	non-dich	false	English	normal chances	Gettierized	side effect	first	11.2 %	0.370	94
		d ("nazi" story Gettierized)	false	true	false	non-dich	true	English	normal chances	Gettierized	side effect	first	30.2 %	1.060	93

	2	a (NEUT: Match1 vs NEG: Match2)	true	true	false	non-dich	true	English	normal chances	Gettierized	other	first	17.3 %	0.570	151
		b (NEUT: Mail1 vs NEG: Mail2)	true	true	false	non-dich	true	English	normal chances	Gettierized	other	third	-4.2 %	-0.130	161
		d (NEUT Clock1 vs NEG Clock2)	true	true	false	non-dich	true	English	normal chances	Gettierized	other	first	5.2 %	0.170	133
Buckwalter 2014	1	(“pump” story Gettierized)	true	true	true	non-dich	true	English	normal chances	Gettierized	side effect	first	30.2 %	1.100	86
	2	(“mayor” story Gettierized)	true	true	true	non-dich	true	English	normal chances	Gettierized	side effect	first	32.3 %	1.320	78
	3	(3rd party “mayor” story Gettierized)	true	true	true	non-dich	true	English	normal chances	Gettierized	side effect	third	17.2 %	0.640	85
Turri 2014	1	a (“normal” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	25.5 %	1.090	42
		b (“Gettier” story)	true	true	true	non-dich	true	English	normal chances	Gettierized	side effect	first	12.5 %	0.420	56
		c (“false” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	44 %	1.670	49
	2	(“no belief” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	40 %	1.280	56
	3	(“false” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	30.2 %	0.980	57
	5	a (“normal” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	31.6 %	1.130	72
		b (“Gettier” story)	true	true	true	non-dich	true	English	normal chances	Gettierized	side effect	first	-4 %	-0.130	78
		c (“false” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	18.3 %	0.500	78
Beebe 2016	2	a (3rd party “environment” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	third	4.3 %	0.150	71
		b (3rd party “movies” story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	third	-0.2 %	0.000	72
		c (3rd party “sales” story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	third	-10.2 %	-0.380	60
		d (3rd party “nazi” story)	false	true	false	non-dich	true	English	normal chances	not Gettierized	side effect	third	11.2 %	0.460	70
	3a	a (3rd party “mayor” story, strong emotional)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	third	6.7 %	0.260	79
		b (3rd party “mayor” story, no emotional)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	third	15 %	0.700	80
	3b	a (3rd party “mayor” story, no observer)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	third	26.7 %	0.820	80
		b (3rd party “mayor” story: strong)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	third	23.3 %	0.800	80
		c (3rd party “mayor” story, none)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	third	18.3 %	0.620	79
	7	a (no outcome info “environment” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	6.5 %	0.180	87
		b (no outcome info “movies” story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	-0.8 %	-0.030	155
		c (no outcome info “sales” story)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	16 %	0.510	87
Wilkenfeld and Lombrozo 2020	1	a (conventional + mechanism)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	7.3 %	0.330	55
		b (conventional + nonmechanism)	false	false	false	non-dich	false	English	normal chances	not Gettierized	side effect	first	5.7 %	0.220	51
		c (moral + mechanism)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	2.3 %	0.230	50
		d (moral + nonmechanism)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	16.2 %	0.670	53

(*Continued*)

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

(*Continued.*)

Paper	Study	Subgroup	Violated	Present	Salient	Scale	Q Phrasing	Language	Chances	Gettierization	Effect	Subject	% difference	d	N
Paprzycka-Hausman 2020	1	a (shooting)	true	true	true	non-dich	true	English	low chances	not Gettierized	other	first	24.7 %	0.740	70
		b (dice – killing/winning)	true	true	true	non-dich	true	English	low chances	not Gettierized	other	first	41 %	1.320	72
		c (dice – throwing six)	true	true	false	non-dich	true	English	low chances	not Gettierized	other	first	22 %	0.730	73
	2	a (shooting)	true	true	true	dich.	true	English	low chances	not Gettierized	other	first	26.5 %	1.286	158
		b (dice – killing/winning)	true	true	true	dich.	true	English	low chances	not Gettierized	other	first	39.8 %	1.464	148
		c (dice – throwing six)	true	true	false	dich.	true	English	low chances	not Gettierized	other	first	8.7 %	1.176	166
	3	(dice – killing/winning)	true	true	true	dich.	true	English	low chances	not Gettierized	other	first	56.2 %	2.545	156
Paprzycka-Hausman 2021*a*	3	a (dice: neutral vs. immoral)	true	true	true	dich.	true	English	low chances	not Gettierized	other	first	22.5 %	0.509	155
		b (cards: neutral vs. immoral)	true	true	true	dich.	true	English	low chances	not Gettierized	other	first	13.7 %	0.316	156
	4	a (slight chance)	true	true	true	dich.	true	English	low chances	not Gettierized	side effect	first	26.7 %	0.733	149
		b (strong chance)	true	true	true	dich.	true	English	normal chances	not Gettierized	side effect	first	8.7 %	0.461	147
	5	a (1/100 chance)	true	true	true	dich.	true	English	low chances	not Gettierized	side effect	first	5.1 %	0.136	160
		b (99/100 chance)	true	true	true	dich.	true	English	normal chances	not Gettierized	side effect	first	6.3 %	0.375	160
	6	a (1/100 chance)	true	true	true	non-dich	true	English	low chances	not Gettierized	side effect	first	26 %	0.830	122
		b (99/100 chance)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	20.3 %	0.730	122
Paprzycka-Hausman 2021*b*	1	("environment" story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	27.3 %	1.020	135
Paprzycka-Hausman 2021*c*	2	(HELP: Mail5 vs HARM: Mail4)	true	true	false	non-dich	true	English	normal chances	Gettierized	other	first	22 %	0.460	139
Yuan and Kim 2021	2a	a ("pump" story Gettierized; Korean)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	first	28 %	0.770	260
		b ("mayor" story Gettierized; Korean)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	first	21.4 %	0.640	258
		c (3rd party "mayor" story Gettierized; Korean)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	third	11.7 %	0.380	235
	2b	a ("pump" story Gettierized; Mandarin)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	first	7.1 %	0.190	299
		b ("mayor" story Gettierized; Mandarin)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	first	20.4 %	0.520	254
		c (3rd party "mayor" story Gettierized; Mandarin)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	third	0.1 %	0.000	277

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

	2c	a (3rd party “pump” story Gettierized; Mandarin)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	third	20.1 %	0.550	138
		b (3rd party “air” story Gettierized; Mandarin)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	third	15.5 %	0.410	147
		c (3rd party “mayor” story Gettierized; Mandarin)	true	true	true	non-dich	true	other	normal chances	Gettierized	side effect	third	-4 %	-0.110	149
Paprzycka-Hausman *et al.* Ms	3	a (replication of Beebe and Buckwalter 2010)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	11.6 %	0.570	126
		b (replication of Beebe and Jensen 2012 [dich. scale])	true	true	true	dich.	true	English	normal chances	not Gettierized	side effect	first	17.8 %	1.632	146
	4	a (low chances “virus” story)	true	true	true	non-dich	true	English	low chances	not Gettierized	side effect	first	38.3 %	1.280	138
		b (normal chances “virus” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	19.3 %	0.750	145
		c (low chances “environment” story)	true	true	true	non-dich	true	English	low chances	not Gettierized	side effect	first	36.7 %	1.200	143
		d (normal chances “environment” story)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	8.9 %	0.350	136
Ryszkowska *et al.* Ms	1	a (group agent [= Members of Board] “movies” story)	false	false	false	non-dich	false	other	normal chances	not Gettierized	side effect	first	31.8 %	1.110	125
Ryszkowska *et al.* Ms	1	b (group agent [= Board] “movies” story)	false	false	false	non-dich	false	other	normal chances	not Gettierized	side effect	first	11.3 %	0.400	107
Zaręba Ms	1	a (k. that the agent caused the side-effect)	true	true	true	non-dich	true	other	normal chances	not Gettierized	side effect	first	18 %	0.630	55
		b (k. that the event caused the side-effect)	true	true	true	non-dich	true	other	normal chances	not Gettierized	side effect	first	12.5 %	0.720	65
		c (k. that state of the environment worsened/improved)	true	true	true	non-dich	true	other	normal chances	not Gettierized	side effect	first	14.3 %	0.410	83
	2	a (k. that the agent caused the side-effect)	true	true	true	dich.	true	other	normal chances	not Gettierized	side effect	first	6.9 %	1.093	155
		b (k. that the event caused the side-effect)	true	true	true	dich.	true	other	normal chances	not Gettierized	side effect	first	15.3 %	0.853	136
		c (k. that state of the environment worsened/improved)	true	true	true	dich.	true	other	normal chances	not Gettierized	side effect	first	17.1 %	0.718	144
Ziółkowski *et al.* Ms *a*	2	a (group/dissent = knowledge)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	20.2 %	0.940	329
		b (group/support = knowledge)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	10.5 %	0.570	321
		c (individual/dissent = knowledge)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	25.2 %	1.070	329
		d (individual/support = knowledge)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	12 %	0.610	321
Ziółkowski *et al.* Ms *b*	4	(knew: fictional corporation/ environment)	true	true	true	non-dich	true	English	normal chances	not Gettierized	side effect	first	10.6 %	0.510	289

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Appendix C. List of all excluded studies

Some studies were excluded because they failed to satisfy one of three criteria:

[1] the design must include a negative and a positive (or neutral) condition
[2] the participants must be asked about knowledge
[3] the experiment must have a between-subject design.
[4] Some studies were excluded for other reasons explained in the following table.

APA reference	Exclusion code and reason	
Beebe J.R. and Jensen M. (2012)	[1]	Study 3a (mayor Emily Spires): the side effect was morally negative in both experimental conditions ("Negative character" and "Positive character").
	[2]	Study 4: the participants were only asked to estimate the probabilities of bringing about the side-effect in the Help/Harm conditions.
	[1]	Study 5: the chances of bringing about the side effect were mismatched in the negative/positive conditions (there were only two groups: slight-chance Harm and strong-chance Help).
Beebe J.R. (2013)	[2]	Study 1: participants were asked only about the protagonist's beliefs.
	[2]	Study 2: participants were only asked about the protagonist's degree of confidence and not knowledge.
	[2]	Study 3: participants were only asked about the protagonist's degree of rational belief.
Beebe J.R. and Shea J. (2013)	[1]	Study 2c (Politician): there was a morally negative event (assassination) in both experimental conditions ("POLITICIAN1" and "POLITICIAN2").
Turri J. (2014)	[1]	Study 4: Only the "Help" condition was tested.
Beebe J.R. (2016)	[2]	Study 1: participants were only asked how much praise/blame the protagonist deserved.
	[2]	Study 4: participants were only asked about the blameworthiness of the protagonists.
	[1]	Study 5: the side effect was negative in both experimental conditions ("Don't care" and "Regret").
	[1]	Study 6: the side effect was negative in both experimental conditions ("Increasing" and "Decreasing").
Wilkenfeld D.A. and Lombrozo T. (2020)	[2]	Study 2: participants were asked four questions (about understanding, explanation) but not about knowledge.
	[2]	Study 3: participants were asked five questions (about understanding, explanation) but not about knowledge.

(*Continued*)

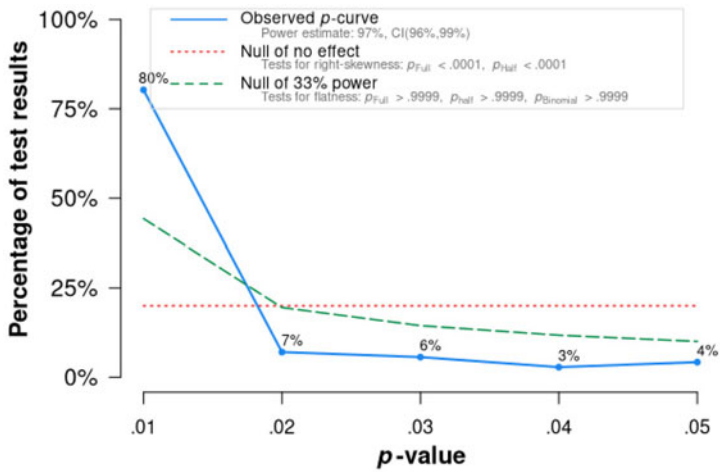
https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

(*Continued.*)

APA reference		Exclusion code and reason
Paprzycka-Hausman K. (2020)	[4]	Study 4: reports the same data as Study 4 from Paprzycka-Hausman 2021*a*, which was included.
Paprzycka-Hausman K. (2021*a*)	[4]	Studies 1, 2 and 3 were excluded because they report the same data as Study 1, 2 and 3 from Paprzycka-Hausman 2020, which were included.
Paprzycka-Hausman K. (2021*b*)	[2]	Study 2: participants were asked only about their attribution of belief.
	[4]	Study 3: introduced too many changes to the experimental design (in particular, it asked about the attribution of knowledge in the explicitly stated absence of belief).
Paprzycka-Hausman K. (2021*c*)	[1]	Study 1: was a pilot study where only a "neutral" condition was tested.
	[1]	Study 3: was a pilot study where only a "neutral" condition was tested.
Yuan Y. and Kim M. (2021)	[1]	Study 1 (Lotto, Zoo, Farm) was not concerned with ESEE.
	[1]	Study 3 (unconfident examinee, prejudiced professor, freaked-out movie-watcher) was not concerned with ESEE.

Appendix D. *p*-curve analysis.

All studies



Appendix Figure 1. *p*-curve plot for all studies displaying the observed *p*-curve and significance results for the right-skewness and flatness test.
Note: The observed *p*-curve includes 71 statistically significant ($p < 0.5$) results, of which 64 are $p < 0.025$. There were 27 additional results entered but excluded from *p*-curve because they were $p > 0.05$.

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

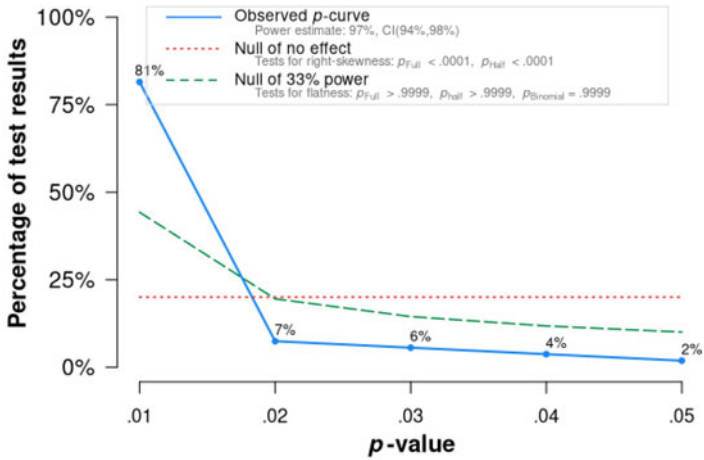
Results:
Right-skewness test:

- full curve (p <0.5): $z = -22.959$, p <0.001
- half curve (p <0.025): z = −22.033, p <0.001

Flatness test:

- full curve (p <0.5): z = 15.771, p >0.999
- half curve (p <0.025): z = 20.634, p >0.999

Power Estimate (95% CI): 97% (95.9–98.5%).



Appendix Figure 2. *p*-curve plot for published studies displaying the observed *p*-curve and significance results for the right-skewness and flatness test.
Note: The observed *p*-curve includes 54 statistically significant ($p < 0.5$) results, of which 49 are $p < 0.025$. There were 25 additional results entered but excluded from *p*-curve because they were $p > 0.05$.

Only published studies

Results:
Right-skewness test:

- full curve (p <0.5): $z = -19.237$, p <0.001
- half curve (p <0.025): z = −18.239, p <0.001

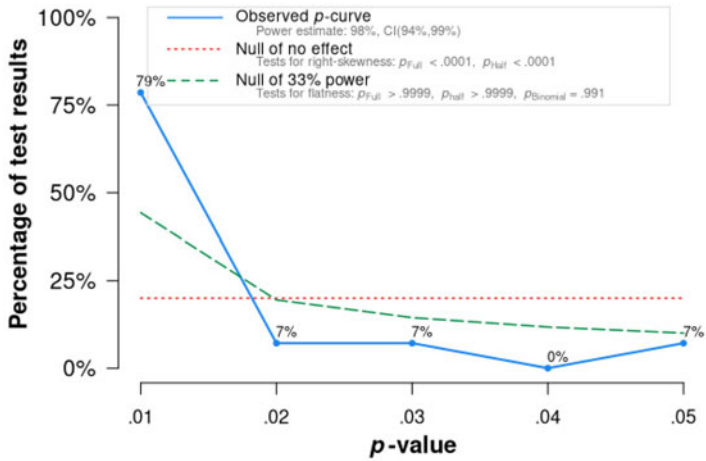
Flatness test:

- full curve (p <0.5): z = 12.946, p >0.999
- half curve (p <0.025): z = 16.976, p >0.999

Power Estimate (95% CI): 97% (94.3–98.1%)

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Close replication studies



Appendix Figure 3. *p*-curve plot for close replication studies displaying the observed *p*-curve and significance results for the right-skewness and flatness test.
Note: The observed *p*-curve includes 14 statistically significant ($p < 0.05$) results, of which 13 are $p < 0.025$. There were 2 additional results entered but excluded from *p*-curve because they were $p > 0.05$.

Results:
Right-skewness test:

- full curve (p <0.5): $z = -9.959$, p <0.001
- half curve (p <0.025): $z = -9.141$, p <0.001

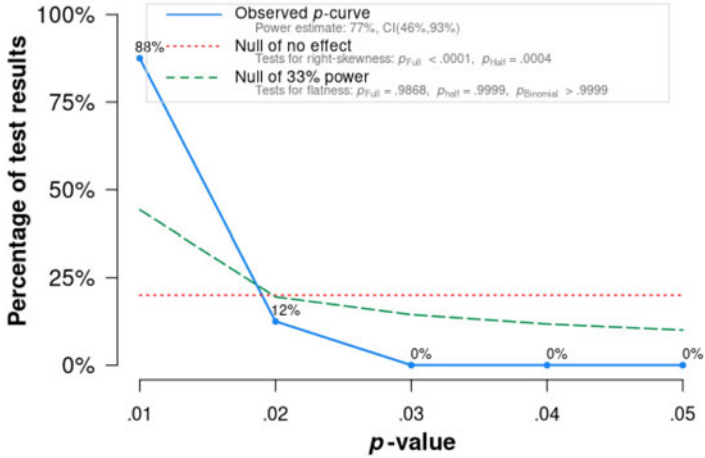
Flatness test:

- full curve (p <0.5): $z = 7.138$, p >0.999
- half curve (p <0.025): $z = 9.344$, p >0.999

Power Estimate (98% CI): 97% (93.7–99%).

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Third-party ESEE studies



Appendix Figure 4. *p*-curve plot for third party studies displaying the observed *p*-curve and significance results for the right-skewness and flatness test.
Note: The observed *p*-curve includes 8 statistically significant ($p < 0.05$) results, of which 8 are $p < 0.025$. There were 8 additional results entered but excluded from *p*-curve because they were $p > 0.05$.

Results:
Right-skewness test:

- full curve (p <0.5): $z = -4.484$, p <0.001
- half curve (p <0.025): $z = -3.324$, p <0.001

Flatness test:

- full curve (p <0.5): $z = 2.221$, $p = 0.987$
- half curve (p <0.025): $z = 3.665$, p >0.999

Power Estimate (95% CI): 77% (46–93%).

Bartosz Maćkiewicz is a research and teaching assistant at the Faculty of Philosophy, University of Warsaw. He is a member of the Laboratory of Experimental Philosophy 'Kognilab'. His research focuses on the metaphilosophical foundations of experimental philosophy, moral reasoning, and philosophy of language. Orcid: 0000-0002-9460-5742, b.mackiewicz@uw.edu.pl

Katarzyna Kuś is an assistant professor at the Faculty of Philosophy, University of Warsaw. She is the head of the Laboratory of Experimental Philosophy 'Kognilab'. Her research focuses on the problem of knowledge, epistemic modals, and the interplay between linguistics and philosophy. Orcid: 0000-0002-1112-766X, kkus@uw.edu.pl

Katarzyna Paprzycka-Hausman is a full professor at the Faculty of Philosophy, the University of Warsaw. She is the head of the Department of Epistemology. Her recent research focuses on philosophy of action (responsibility-based approach to agency, omissions account of the Knobe effect) and experimental philosophy (consequence account of the epistemic side-effect effect). Orcid: 0000-0001-9956-6650, kpaprzycka@uw.edu.pl

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press

Marta Zaręba is an assistant professor at the Faculty of Philosophy, University of Warsaw. Her research focuses on philosophy of action, with a special interest in issues related to the role of intentions. Orcid: 0000-0003-0741-062X, zareba.ma@uw.edu.pl

Cite this article: Maćkiewicz B, Kuś K, Paprzycka-Hausman K, Zaręba M (2022). Epistemic Side-Effect Effect: A Meta-Analysis. *Episteme* 1–35. https://doi.org/10.1017/epi.2022.21

https://doi.org/10.1017/epi.2022.21 Published online by Cambridge University Press