

1-1-2016

# A Systematic Review of Studies Using Gettier-Type Thought Experiments

Maximilian Popiel

*University of Denver*, [mpopiel@gmx.com](mailto:mpopiel@gmx.com)

Follow this and additional works at: <http://digitalcommons.du.edu/etd>

---

## Recommended Citation

Popiel, Maximilian, "A Systematic Review of Studies Using Gettier-Type Thought Experiments" (2016). *Electronic Theses and Dissertations*. 1216.

<http://digitalcommons.du.edu/etd/1216>

This Thesis is brought to you for free and open access by the Graduate Studies at Digital Commons @ DU. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of Digital Commons @ DU. For more information, please contact [jennifer.cox@du.edu](mailto:jennifer.cox@du.edu).

A Systematic Review of Studies Using Gettier-Type Thought Experiments

A Thesis

Presented to

the Faculty of the Morgridge College of Education

University of Denver

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

by

Maximilian Popiel

November 2016

Advisor: Kathy Green

Author: Maximilian Popiel

Title: A Systematic Review of Studies Using Gettier-Type Thought Experiments

Advisor: Kathy Green

Degree Date: November 2016

### **Abstract**

Experimental epistemologists have recently begun using Gettier-type thought experiments to test various philosophic theories. Generally speaking, these thought experiments present a justified, true belief which intuitively does not seem like knowledge. Despite the studies using this same general definition, they have exercised a myriad of different particular Gettier cases and experimental methods. Some results have been conflicting, or otherwise counter-intuitive, and interpretations of their findings have been divergent. The present study was a systematic review of these experiments, with a focus on experimental methods. Studies were compared on readability and factors effecting participant fatigue and comprehension. The results suggest that readability scores and potential for fatigue were inconstant across included studies. Readability scores varied from elementary to college reading level; similarly, some studies presented only a single case to participants, and others presented as many as nine. Ultimately, the present study makes recommendations to future researchers using Gettier-type thought experiments, in hope that better empirical rigor will improve consistency of experimental findings.

## Table of Contents

|  |     |
|--|-----|
| Abstract.....                              | ii  |
| Table of Contents .....                    | iii |
| Introduction.....                          | 1   |
| Literature Review.....                     | 4   |
| Other Studies Using Gettier Cases.....     | 9   |
| Comprehension Questions.....               | 11  |
| Reading Level.....                         | 12  |
| Problem Statement.....                     | 14  |
| Method.....                                | 15  |
| Search Strategy.....                       | 16  |
| Inclusion and Exclusion Criteria.....      | 18  |
| Search Results.....                        | 18  |
| Table 1.....                               | 20  |
| Results.....                               | 21  |
| Table 2.....                               | 22  |
| Study Motivations and General Methods..... | 23  |
| Readability Analysis .....                 | 26  |
| Participant Fatigue .....                  | 28  |
| Comprehension Questions.....               | 30  |
| Discussion.....                            | 32  |
| References.....                            | 38  |
| Appendix A.....                            | 42  |
| Appendix B.....                            | 43  |

## **Introduction**

“Gettier case” is an umbrella term for any instance where a justified, true belief does not intuitively seem to qualify as knowledge. These cases are often presented in the form of thought experiments and recently experimental epistemologists have begun testing them empirically. However, results have been conflicting, and a myriad of different techniques have been employed in the studies. For instance, there does not appear to be a consensus about the use of comprehension questions, nor attention paid to the reading level of the thought experiments. Other empirical issues remain besides these, such as infrequent reporting of effect size, lack of reliability analysis, and the potential for participant fatigue. The present study was a systematic review of empirical studies using the Gettier-type thought experiments. The focus of the review was on study characteristics that, if controlled, may lead to improvement in empirical rigor in experimental use of thought experiments and thus reduce the number of findings that are not replicated.

One use of the Gettier cases has been to test cross-cultural differences on epistemic intuitions. However, these studies have been criticized for employing small sample sizes and their results conflict (e.g., compare Weinberg, Nichols, and Stich, 2001, and Machery et al., 2015, to Seyedsayamdost, 2015, and Kim and Yuan, 2014). In

addition to these differences in reports of cross-cultural group differences on Gettier intuitions, there are also conflicting reports about whether there is a gender effect on epistemic intuitions. In a review of studies testing gender differences in philosophical intuitions, at least three studies with conflicting findings were identified using Gettier-type thought experiments (Adleberg, Thompson, & Nahmias 2014). Other studies have entirely different aims such as generating epistemic theory (Powell, Horne, Pinillos, & Holyoak, 2015), or comparing folk intuitions with those of philosophers (Starmans & Friedman, 2012).

One point that seems somewhat absent from the literature is whether participants actually understand the thought experiments properly. This concern is highlighted by the fact that many of the cross-cultural studies give these thought experiments to individuals for whom English may not be a first language. To remedy this, some studies employ the use of comprehension questions asking a basic fact about an important aspect of the presented Gettier case. Unfortunately, not all studies even report how many participants are actually excluded by answering these questions incorrectly, so their efficacy is unclear. However, even a single, yes/no comprehension question might increase data quality over no comprehension question. Given the difficulty of reproducing results in this field, their use seems warranted.

Finally, there is evidence that Gettier cases are sensitive to order effects (Swain, Alexander, & Weinberg, 2008) and small changes in wording or participant instructions

can drastically alter the results (Cullen, 2013). It is also not clear that studies using Gettier cases have accounted for these findings.

For the present review, basic study information was compiled. This information included hypotheses, sample size, measures employed (if reported), and results. The number of thought experiments given to participants was also recorded. Anecdotally, these thought experiments seem to be mentally taxing for respondents. There does not appear to be any research on participant fatigue in this specific area and the problem ostensibly becomes more pronounced as the number of thought experiments employed increases. Moreover, if we take seriously concerns of the sort such that participants' intuitions "require at least a modicum of reflection" (Weinberg, Nichols, & Stich, 2001, p. 450), then we should be particularly worried that fatigued respondents will not be providing the type of intuitions we require. If the studies employ multiple thought experiments administered to the same participant at once, some form of counterbalancing should be employed, as Gettier-type thought experiments are apparently sensitive to order effects (Swain, Alexander, & Weinberg, 2008). Where applicable, it will be noted whether the study counter-balanced or randomized the order of the thought experiments.

The particular thought experiments used were cataloged and a reading level analysis performed, using the Dale-Chall Readability Index (Chall & Dale, 1995). The definition of a Gettier case was exceedingly inclusive for the present analysis, if the author(s) identified their thought experiment as a Gettier case, it was considered as such. In addition, "control" cases and similar thought experiments (such as the skeptical

pressure cases) were included if the study used them in conjunction with “true” Gettier cases (again, as indicated by the author). The reason for this is that it would be beyond the scope of this review to take a philosophical position as to what does or does not qualify as a Gettier case, given that the present motivation was to broadly understand how these thought experiments are being used experimentally.

The present review likely serves not only experimenters working with Gettier cases, but those using thought experiments as any type of empirical instrument. An emphasis on experimental rigor can increase reliability of results and better guide future research. Since thought experiments are somewhat unfamiliar to empirical researchers there does not appear to be substantial work on how to best use them. This work aims to move towards a better understanding of how Gettier cases, and thought experiments more broadly, are and should be used. Accordingly, the purpose of this study was to better understand how Gettier cases are used experimentally and provide insight into why so many study results have been conflicting



## Literature Review

When is belief considered knowledge? This is a question that has interested philosophers and scientists for thousands of years. To any knowledge generating practice it seems utterly foundational; however it has been difficult to answer definitively. One idea that has seen considerable support is the *Tripartite Analysis of Knowledge* that holds that a person, or  $S$ , holds knowledge of  $p$  if (and only if)  $p$  is true,  $S$  believes  $p$ , and  $S$  is a justified belief (Stanford Encyclopedia of Philosophy, 2001). Thus, to the original question, this analysis of knowledge implies that knowledge is having a justified, true belief (JTB).

To understand what it means to be true and justified, consider, for example, the Pythagorean Theorem. We might consider our belief in it justified since we were taught the theorem in school, and it seems true since it holds for every right triangle we have tested. Or, if it is shown by way of deductive proof, then we can jointly satisfy its truth and justification conditions. Withholding debate over satisfying its conditions, the theory makes intuitive sense. Conversely, there are a number of JTBs that intuitively do not seem to be knowledge. This was brought to light by Edmund Gettier in an influential paper where he presented a series of these examples. Since then philosophers have come

up with numerous Gettier cases, with agents that have a JTB that intuitively does not seem to qualify as knowledge (Gettier, 1963). More recently, a Gettier case has come to refer to essentially any JTB that intuitively does not seem like knowledge (Cullen, 2013).

One of the prototypical Gettier cases (similar to, although not exactly one Gettier himself penned) is the “car case” presented in Appendix A. In this case, Bob seems justified in his belief that Jill has an American car, and indeed this belief is true. However when asked if this Bob “really knows” that Jill drives an American car, the majority of people decline to attribute knowledge. (This is at least true of Western respondents; see Weinberg, Nichols, & Stich 2001; this paper is discussed in greater depth later in the present review.)

For a very different example, consider also the “fake barn” case (again see Appendix A). Here, the agent (Henry) holds the JTB that he is looking at a barn, but he just as easily could be looking at a fake. Perhaps we do not want to deny him knowledge at present, but we would be inclined to in the latter case. The problem here is that from his standpoint there is no difference – his knowledge falls on mere “epistemic luck.” One might argue that this is fine, and that indeed we might never know the absolute truth of certain propositions; however, then, a JTB would no longer be sufficient for knowledge. Knowledge would now require a JTB and the appropriate epistemic luck. This is not a satisfactory position: it leaves open the idea of whether we ever actually have knowledge at all, as we could never know where our epistemic luck would fall.

These Gettier cases certainly seem to call into question whether a JTB is sufficient for knowledge – or do they? Perhaps we should not simply rely on intuitions as proof; they might be pliable or inconsistent. Or we might be putting the cart before the horse whereas our thought experiments (here we mean Gettier cases) might not even be accessing the type of intuitions we think they are. Or perhaps these intuitions are not as widely shared, as some philosophers believe (see Turri, 2016). In any case these are now empirical questions: are thought experiments reliable instruments to test intuitions, and are the (apparent) intuitions themselves universal?

Further, we must consider whether the epistemic processes we wish to illuminate are normative or descriptive practices. The former is arguably beyond the reach of empirical study as, by its nature, empiricism assumes certain normative axioms about knowledge. But, the descriptive project seems ripe to be informed by social science research. And while a number of non-empirical questions remain (e.g., are Gettier intuitions meaningful such that they would be relevant or interesting to philosophers?), experimental philosophers have tested these cases to tease out what they can. One current idea is that if there is a systematic difference among peoples' intuitions in this area, philosophers might be wrong to deny knowledge to the Gettier agents (Turri, 2016). As to the normative implications, perhaps differing intuitions suggest that constructing knowledge (ostensibly a universal concept) wholly as JTBS results in a form of epistemic chauvinism.

Perhaps the most notable study testing group differences with Gettier cases is Weinberg, Nichols, and Stich's 2001 paper *Normativity and Epistemic Intuitions*. They

tested whether Western students and Indian Sub-Continent (SC) or Southeast Asian (EA) students differed in rates of knowledge ascription. This work was largely inspired by that of Nisbett, Peng, Choi, and Norenzayan (2001) who, drawing on a myriad of evidentiary sources, suggested that Westerners had more analytic cognitive styles, while EAs were more collectivist. While the latter paper argued from varied sources of (mostly anecdotal) evidence, it did not include any mention of the Gettier cases and, notably, did not mention the cognitive styles of SCs. They were included in the Weinberg, Nichols, and Stich (2001) study for largely for idiosyncratic reasons. (The paper states it “had no theoretical basis for expecting that the epistemic intuitions of people from the Indian sub-continent [would vary],” p. 14).

The Weinberg et al. (2001) paper ultimately used 7 thought experiments across 11 studies, 3 of which are unique to that paper. (The “special feeling” case would probably not be considered a Gettier case by the current definition, and the authors make no claim that it is, as it is not clear that a future conditional statement could be a JTB. However, the thought experiment is similar enough to the Gettier cases for the present discussion.) Of these, 7 had statistically significant differences in rates of knowledge attribution (at  $p = 0.05$ ). The sample size for these studies ranged from 41 to 237 with a median sample size of 86 (Weinberg, Nichols, & Stich 2001). Notably, their EA and SC samples were particularly small in comparison to their Western sample with median proportions of 0.13 and 0.35 (for EA/SC to Western participants, respectively). For their investigation of SES the proportion of low SES to high SES was 0.71.

Perhaps that is why Weinberg et al.'s (2001) results have not been replicated. Seyedsayamdost (2015) criticized their study for having a small sample size, and conducted a replication of their comparative EA and SC trials. His sample sizes were somewhat larger (median sample size was 111 compared to 86 for the Weinberg et al. study), and the median proportion of EA to Western respondents was also more balanced at 0.48 and at 0.65 for SC participants. He was unable to replicate the findings of the Weinberg paper. Similarly Kim and Yuan (2015) conducted a replication with an even larger EA sample (the proportion of EA to Western participants was 2.83:1, overall  $n = 140$ ) and also found no statistically significant difference.

In the concluding section of their paper, Weinberg et al. (2001) respond to the objection that their effect size is small and “philosophically disinteresting” by arguing, essentially, that the effect was neither small nor disinteresting: “[the effect sizes] are quite comparable... social psychologists take to show important differences between groups” (p 448). Still, regarding the latter concern, perhaps the issue is both that their effect size was small in conjunction with their sample size being small. While they did not actually report their effect sizes, the median odds ratio seemed to be 0.53, with the largest being when they tested the Zebra-in-zoo case between high and low SES respondents at 3.75; while this is certainly not a small effect size, less attention seems to have been paid to the SES group comparisons. Weinberg and colleagues consider their “second reply [to be] more important” (Weinberg et al., 2001, p. 448), a reply that speaks more to that latter concern (i.e., that these differences are philosophically disinteresting). Again, this concern is beyond the scope of the present paper, and better left to philosophical debate.

At this point, one might be inclined to write off the results of Weinberg et al. (2001) as an anomaly. However, they are not the only study to report cross-cultural differences in Gettier cases. For example, Machery et al. (2015) published a study where Gettier cases were given to participants in Brazil, Japan, India, and the USA and found a number of differences. Notably, Indian participants were significantly less likely to deny knowledge than the other three samples, but Japanese participants were not. Interestingly, these differences disappeared on the second knowledge probe which asked if *S* really knew *p*, or if *S* “feels like [s]he knows [p] but [s]he doesn’t actually know [this]” (Machery et al., p. 5). Still, this study found little evidence to support the claim that EA (or at least Japanese) participants had different intuitions on the Gettier cases. Since this result has failed replication in at least two other studies (Kim & Yuan, 2014; Seyedsayamdost, 2015) it is probable that this finding was indeed aberrant.

### **Other Studies Using Gettier Cases**

In addition to racial/ethnic differences on Gettier intuitions, there are also conflicting reports about whether there is a gender difference in epistemic intuitions in general. In a review of studies testing gender differences in philosophical intuitions, three studies were identified using Gettier-type thought experiments to assess a gender difference, one of which was an unpublished meta-analysis by John Turri. While the individual studies did not find a gender-effect, Turri’s meta-analysis did (as cited by Adleberg, Thompson, & Nahmias, 2014, p. 633). Unfortunately, the meta-analysis was unpublished and the data are no longer available for review (J. Turri, personal communication, May 21, 2016). When the Adleberg et al. review presented a Gettier

case to 136 students and there was indeed a difference between how men and women responded, the difference was not statistically significant (Adeberg, Thompson, & Nahmias, 2014). In addition, most (if not all) of the aforementioned studies looking at cross-cultural differences in intuitions - also tested to see if there was a gender-effect (largely on grounds of experimental rigor) and did not report significant differences. While Adeberg, Thompson, and Nahmias (2014) make a convincing case for a gender-effect in philosophical intuitions, such differences have not been made apparent through the use of Gettier cases.

Still other researchers have tested specific conditions of the Gettier cases themselves. A prime example of this is found in Nagel, Juan, and Mar's 2013 paper *Lay denial of knowledge for justified true beliefs*. Their idea was to present Gettier cases along with "standard true belief," false belief, and "skeptical pressure" variants. These variants changed the conditions of the case slightly such that the agent would not have a JTB that intuitively does not seem like knowledge (as in the standard Gettier case), but rather they would have a true (or false) belief. The skeptical pressure cases are slightly different; they present an agent with a JTB but simply imply that the agent might be mistaken. The fake barn case readily makes itself into a skeptical pressure case – we could simply imply that, even if Henry is driving in an area without any fake barns, he would not know whether or not any barn he looked at was real or fake and we might therefore be inclined to deny him knowledge. (Notice how taken to its conclusion here, the skeptical pressure would seem to ultimately seem to deny all knowledge, hence its name.)

Even though one could introduce this skeptical pressure to almost any claim of knowledge, simply mentioning it seems to encourage participants to deny knowledge to an agent. Nagel et al. (2013) found that only 41.5% of participants attributed “unwavering” knowledge to the agent in the skeptical pressure cases (this compared to 75.3% in the standard true belief condition and 14.3% in the false belief condition). It is worth noting that Nagel and her collaborators had three possible responses for these cases: immediate knowledge denial, unwavering knowledge ascription, and delayed knowledge denial. This is because participants were given three options for the “does S really know P” question – “Yes”, “No”, and “Unclear.” If participants chose “Yes”, they were probed further as to whether S “really knew” P or S “feels like they know P, but doesn’t actually know it.” The former was coded as “unwavering knowledge attribution” and the latter as “immediate denial.

It seems that the unwavering attribution and immediate denial can be interpreted more or less as the knowledge attribution or denial in the previously discussed dualistic cases, while the delayed denial can complicate direct comparison of this study to the others. One plausible course of action is to ignore the middle category entirely, despite its likely philosophical interest, artificially reducing the study to having had only two response categories.

### **Comprehension Questions**

One point we have yet to consider is whether participants actually understand the thought experiments properly. This concern is highlighted by the fact that many of the cross-cultural studies give these thought experiments to individuals for whom English



may not be a first language. To remedy this, some studies employ the use of comprehension questions. When employed they are often yes/no questions asking about some basic tenet of the thought experiment. For example, in the car case it might ask: does Jill drive an American car? And participants would have to answer “yes” to proceed with the survey.

Unfortunately, some studies did not report how many participants are actually excluded by answering these questions incorrectly, and even fewer reported if the exclusion of participants by comprehension questions affected their results. The efficacy of the questions is, admittedly, unclear. It is worth noting that a yes/no comprehension question would still allow half of the participants to continue if they were randomly guessing, without having read the thought experiment at all. However, it seems unlikely that their inclusion would negatively affect data quality. At the very least, including a set of comprehension question will give the researcher the option to remove participants who fail said question(s) if the need arises.

### **Reading level**

It seems there are almost as many Gettier cases as there are studies employing them. How do we know these are all comparable? From a philosophical point of view, any collection of sentences that present a JTB that does not seem like knowledge might qualify. However these studies are administering the thought experiments to laypeople, often college students, but sometimes individuals with little or no college education whatsoever. We cannot assume that the reading difficulty of these thought experiments is equal.

The Dale-Chall Readability Index (Chall & Dale, 1995) was developed to help measure readability, with scores indicative of the years of formal education required to understand the passage (e.g. a Dale-Chall score of 5- 5.9 would indicate a fifth to sixth grade reading level). It is normed against words known by US fourth grade students (Ibid). Perhaps as a result on the formula's evidentiary basis, it has been found to perform marginally better than other readability formulas in detecting problematic survey questions (Lenzner, 2014). The cancer Conspiracy Case, for example, has a Dale-Chall score of 9.6 while the Buick has a Dale-Chall score of only 6.1 (see Appendix B). Especially if we wish to administer these cases to individuals for whom English is not a first language, this is an issue that could contribute to Type I error. Weinberg et al. (2001) used education as a proxy measurement for SES, and found that low SES individuals (or at least, those with less education) were more likely to attribute knowledge in the cancer case. This is another result that has not been replicated (Seyedsayamdost, 2015). Perhaps this difference was due to the reading difficulty of the cancer conspiracy case. If experimenters wish to continue creating Gettier cases, it would be best if the reading levels were comparable, or disparities in reading level acknowledged and potentially treated as a covariate.

## **Problem Statement**

Although Gettier cases have been used to test a variety of hypotheses, the results of the studies conflict. There is little uniformity in the way Gettier cases are utilized experimentally, except when the studies are pure replications. Since Gettier cases, and thought experiments more generally, are a somewhat novel empirical tool, little is known about their reliability as a scientific instrument. To help understand why their use has led to so many conflicting findings, the present study is a systematic review with an emphasis on experimental method. The guiding direction was that idiosyncratic study differences have led to a divergence of experimental findings, and accordingly the findings of these studies may not be comparable to one another.

## **Method**

A systematic review was conducted of empirical studies using the Gettier-type thought experiments. Basic study information was compiled such as hypotheses, sample size, reading level, and psychometric quality of measures employed (if reported).

The number of thought experiments given to participants was also recorded. Anecdotally, these thought experiments seem to be mentally taxing for respondents. Although there appear to be no data on participant fatigue from Gettier cases, this problem ostensibly becomes more pronounced as the number of thought experiments employed increases. Moreover if we take seriously concerns of the sort such that participants' intuitions "require at least a modicum of reflection" (Weinberg, Nichols, & Stich, 2001, p. 450), then we should be particularly worried that fatigued respondents will not be giving the type of intuitions we require. If the studies employ multiple thought experiments administered to the same participant consecutively, some form of counterbalancing should be employed, as Gettier-type thought experiments are apparently quite sensitive to order effects (Swain, Alexander, & Weinberg, 2008). Where applicable, it was noted whether or not the study counter-balanced or randomized the order of the thought experiments.

The particular thought experiments used were cataloged and a reading level analysis was performed, using the Dale–Chall readability formula. The definition for a Gettier case was over-inclusive for the present analysis, if the author(s) identify their thought experiment as a Gettier case, it was considered as such. In addition, “control” cases and similar thought experiments (such as the skeptical pressure cases) was included if the study used them in conjunction with “true” Gettier cases (again, as indicated by the author). The reason for this is that it would be beyond the scope of this review to debate what does or does not qualify as a Gettier case, given that the present motivation is to broadly understand how these thought experiments are being used experimentally. If the study compared Gettier cases to skeptical pressure cases or other cases, it seemed appropriate to analyze the reading levels of both.

In addition to the reading level analysis, studies were compared on their use of a comprehension question (or lack thereof). The response categories were noted, as was the number of participants excluded. Most studies use yes/no comprehension questions (for example, Machery et al., 2015), which are probably not optimal for reasons previously discussed.

### **Search strategy**

A Boolean search phrase was entered into Philosopher's Index, Cambridge Companions Online: Philosophy, Religion, and Culture, Proquest (dissertations and theses), Psychinfo, Academic Search Complete, PubMed, PLOS One, Philpapers.org, and Google Scholar (excluding patents). The Boolean phrase used was:

("folk epist\*" OR "holistic epist\*" OR "analytic epist\*" OR knowledge OR cognit\* OR inferen\* OR intuition\*) AND (gettier\*) AND ("experimental philosophy" OR survey OR study\* OR experimen\*)

The strategy could be expanded by including "Kripke" as a keyword, as some studies use Kripke-esque, as opposed to Gettier, thought experiments (e.g., Beebe, Undercoffer, in press). However, it is unclear if these studies are similar enough in scope and/or methods for the present analysis and so this was not done in the current study.

Abstracts of papers yielded by the academic databases by were screened for relevance, any paper that did not mention having an empirical study or survey in the abstract was excluded. Both backward and forward search strategies were employed on all studies not excluded by abstract screening. Google Scholar was used for forward searching. In addition, theoretical (philosophical) papers discussing the topic of empirically testing thought experiments were also back-searched. Finally, papers that provided initial motivation for studies of this type were forward searched (e.g., Nisbett, Peng, Choi, & Norenzayan, 2001). The same method of abstract screening was used with papers recovered by the forward and backward searches.

To help counteract publication bias, unpublished dissertations and theses were included in the database search. In addition, the personal websites of researchers doing work in this area were searched, as were repositories of relevant research groups (e.g., the Experimental Epistemology Research Group based in SUNY Buffalo). Again these studies were screened in the same manner as the published papers.

## **Inclusion and Exclusion Criteria**

Any study that suggests its thought experiments are “Gettier cases” was considered for inclusion as long as it was in English, accessible in full-text, published between 2001 and 2016, and included empirical testing of Gettier thought experiments. Studies that did not use Gettier-type thought experiments, or did not present original data, were excluded. Qualitative studies were also excluded.

If the studies did not report some information of interest to the present review (such as the number of participants excluded through the comprehension questions), it was not automatically excluded, as removing such studies would have left few, if any, to work with. This is another issue which, if not clear already, the present review intended to highlight: many of the studies in experimental philosophy do not provide as much study information as many researchers have come to expect.

## **Search Results**

Entering the Boolean search phrase in Google Scholar yielded over 4,300 results, so only the first 800 were considered. (Google Scholar orders the search results based on relevancy, after the first hundred or so results it was clear the remaining results would not be useful.) Entering the same in the research databases altogether yielded 1649 potential papers; however these were not all unique hits, as many papers appeared in more than one database. After screening abstracts, there were approximated 35 studies to be considered further. Included studies were written in English, accessible in full-text, published between 2001 and 2016, and included empirical testing of Gettier thought

experiments. Studies that did not provide critical information about their research methods (sample sizes, full text of thought experiments, etc.), tested non-epistemic intuitions (e.g., moral intuitions), or did not use Gettier-type thought experiments were excluded. Qualitative studies were also excluded. After these criteria were implemented, 17 studies remained. Two papers were found not to be presenting original data and were excluded; after this final exclusion, only 15 studies remained. These are presented in Table 1 alphabetically by author.



Table 1

*Author, Year, Study Purpose, Sample Size, Use of Comprehension Questions, and Number of Excluded Participants.*

| Author(s)                            | Publication year | Purpose  | Total sample size | Comprehension question  | Participants excluded     |
|--------------------------------------|------------------|--|-------------------|---|---------------------------|
| Adleberg, Thompson, & Nahmias        | 2015             | Gender effect  | 136               | quality check   | 20                        |
| Buckwalter                           | 2014             | Testing the “epistemic side effect”                                      | 249               | No  | n/a                       |
| Colaço, Buckwalter, Stich, & Machery | 2014             | Intuitional difference between high-low knowledge defeaters <sup>a</sup> | 234               | 2 binary  | 65                        |
| Cullen                               | 2013             | Survey design  | 1082              | No  | n/a                       |
| Kim & Yuan                           | 2014             | Cross-cultural differences   | 202               | No  | n/a                       |
| Machery et al.                       | 2015             | Cross-cultural differences   | 521               | 1 binary per case   | 271                       |
| Nagel, Juan, & Mar                   | 2013             | Comparing Gettier cases to skeptical pressure cases                      | 208               | 2 binary per case   | not reported <sup>b</sup> |
| Powell, Horne, Pinillos, & Holyoak   | 2015             | Construction of epistemic theory   | 2433              | quality check and scoring system  | 481 <sup>c</sup>          |
| Seyedsayamdost                       | 2015             | Cross-cultural & SES differences   | 2098              | No  | n/a                       |
| Starmans & Friedman                  | 2012             | Differences between philosophers and laypeople                           | 417               | 1 or 3 binary   | 84                        |
| Swain, Alexander, & Weinberg         | 2008             | Order effects of multiple cases  | 228               | excluded for “working with a different concept of [knowledge]” <sup>d</sup> | 8                         |
| Turri                                | 2013             | Replications of previous findings  | 506               | At least 2 binary   | 82                        |
| Turri, Buckwalter, & Blouw           | 2015             | Testing the role of epistemic luck                                       | 1665              | 2 binary  | 124                       |
| Weinberg, Nichols, & Stich           | 2001             | Cross-cultural & SES differences   | 1090              | No  | n/a                       |
| Ziółkowski                           | 2015             | Testing the role of epistemic luck                                       | 480               | 2 binary  | 50                        |

<sup>a</sup> Refers to how likely participants were to attribute knowledge in “unproblematic cases of knowledge”; no effect was found and the paper instead reported finding a significant age difference.

<sup>b</sup> Original sample size 238, reports removing 30 participants for not finishing the survey, or finishing it too quickly.

<sup>c</sup> Paper coded response quality under a “liberal” and “strict” system. The former only removed 232 participants, the latter is reported in the table. They also removed 21 participants who failed the quality check question(s).

<sup>d</sup> P. 139, Swain, Alexander, & Weinberg, 2008

## Results

The results of the analyses are split into four general sections: motivations and general methods, readability analysis, participant fatigue, and comprehension questions. Table 1 presents information about study motivations and general methods, such as sample size and comprehension questions. Table 2 presents the particular Gettier cases used in each study and information relating to participant fatigue such as the number of cases given at once and the weighted average readability of the cases used in the study. The full text of each Gettier case and its readability score is presented in Appendix B. The prompt for each Gettier Case was standardized for the readability analysis where each case ended with the prompt “*Does S really know p, or only believe it?*”.

Table 2

*Author, Number of Cases Given, Cases, Readability Score, and Score Range.*

| Authors                              | Cases given <sup>a</sup> | Cases used   | Dale-Chall Score <sup>b</sup> | Range <sup>c</sup> |
|--------------------------------------|--------------------------|--|-------------------------------|--------------------|
| Adleberg, Thompson, & Nahmias        | 7                        | Watch <sup>d</sup>   | 5.70                          | n/a                |
| Buckwalter                           | 1                        | Mayor, Pump  | 6.67                          | 0.2                |
| Colaço, Buckwalter, Stich, & Machery | 1                        | Fake-barn v. 1   | 7.5                           | n/a                |
| Cullen                               | 1-3                      | Broken Clock v. 1, Emmy Award, Truetemp, Buick, Special Feeling  | 6.18                          | 2.6                |
| Kim & Yuan                           | 1                        | Buick  | 6.1                           | n/a                |
| Machery et al.                       | 4                        | Hospital, Las Vegas, Furniture, Diamond v. 1   | 6.65                          | 1.5                |
| Nagel, Juan, & Mar                   | 9                        | Baby Deer, Buick, Car crash, Mall Robbery, Las Vegas, Clock, ID card, Diamond v. 1, Furniture store      | 6.31                          | 2.1                |
| Powell, Horne, Pinillos, & Holyoak   | 1 <sup>e</sup>           | Diamond v. 1, Detective, Conflict Diamond, Vegan Cookie  | 6.66                          | 0.9                |
| Seyedsayamdost                       | 1-4                      | Buick, Cancer Conspiracy, Truetemp, Zebra v. 1   | 6.29                          | 8.1                |
| Starmans & Friedman                  | 1-2                      | Pen Burglar v. 1, Watch Burglar, Yogurt, Coin  | 5.68                          | 1.1                |
| Swain, Alexander, & Weinberg         | 4                        | Special feeling, Truetemp, Fake-barn v. 2, Poison gas  | 6.55                          | 2.6                |
| Turri                                | 1                        | Pen Burglar v. 1, Coin v. 2, Zebra v. 2, Pen Burglar v. 2  | 5.76                          | 0.7                |
| Turri, Buckwalter, & Blouw           | 1                        | Squirrel, Diamond v. 2   | 6.42                          | 0.6                |
| Weinberg, Nichols, & Stich           | 1                        | Truetemp, Truetemp (Community), Truetemp (Elders), Buick, Cancer Conspiracy, Zebra v. 1, Special Feeling | 6.16                          | 8.1                |
| Ziólkowski                           | 1                        | Fake Barn v. 1, Stopped-Clock, Thermometer   | 7.07                          | 0.8                |

<sup>a</sup> Refers to the number of cases given at once to each participant

<sup>b</sup> Weighted average

<sup>c</sup> Difference between highest and lowest readability scores in studies with multiple thought experiments

<sup>d</sup> Although a total of 7 thought experiments were given to participants, only 1 was a Gettier case: it was presented fourth in the series.

<sup>e</sup> Also gave participants a distracter story after the Gettier cases, but before the recall task.

## Study Motivations and General Method

Approximately half of the studies (n=7) were comparative in nature looking to find a group difference on Gettier intuitions. As discussed earlier, multiple studies investigated a cross-cultural difference (n=5; see Table 1), and two of these also investigated differences across SES (Seyedsayamdost, 2015 and Weinberg, Nichols, & Stich, 2001). At least one study explicitly investigated a gender effect (Adleberg, Thompson, & Nahmias, 2015), and one compared philosophers' intuitions to those of laypeople (Starmans & Friedman, 2012). One study (Colaço, Buckwalter, Stich, & Machery, 2014) reported an age effect, where older people were more likely to attribute knowledge, however that finding reads somewhat like a data-mining exercise. Their original purpose was to determine if individuals who were more likely to deny knowledge in other situations would be similarly more likely to do so on Gettier cases, however no effect was found (Ibid).

Of the remaining studies, two tested the role of epistemic luck (Turri, Buckwalter, & Blouw, 2015; Ziólkowski, 2015), one compared Gettier cases to skeptical pressure cases (Nagel, Juan, & Mar, 2013), and one study attempted construction of epistemic theory (Powell, Horne, Pinillos, & Holyoak, 2015). In addition, one study (Buckwalter, 2014) tested the role of the “epistemic side-effect,” wherein intuitions can change if there is a moral component to the Gettier case (e.g., if the agent is morally deficient).

The remaining two studies found something akin to order effects of Gettier cases. Swain, Alexander, and Weinberg (2008) argued in their paper that intuitions can change as a result of the order in which they are presented. Cullen (2013) countered by arguing

that it may not be the actual intuitions that are changing, but simply the way participants report them. He investigated how changing the order, or making minor changes to the survey prompts had fairly drastic results on participants' responses. How one could tease out if the participants' actual intuitions are changing, or if the difference is merely an artifact of survey design is not altogether clear. However his main point is well taken - responses on Gettier cases seem particularly sensitive to minor changes in survey design. And taking these two studies into consideration, it is certainly clear that there are order effects for the Gettier cases and counterbalancing is probably warranted. While some studies report doing this (e.g., Nagel, Juan, & Mar, 2013), at least one reports using a fixed-order (Machery et al., 2015).

About half of the studies (n=7; see Table 1) report using at least 1 comprehension question. All the comprehension questions were reported as binary. Two studies (Adeberg, Thompson, & Nahmias, 2015 and Powell, Horne, Pinillos, & Holyoak, 2015) instead used a quality check question (e.g., "If you are paying attention to this question, select 'Really Knows'"). One study (Swain, Alexander, & Weinberg, 2008) excluded for "working with a different concept of [knowledge]." It is not altogether clear if this served as an improvement to data quality. Finally, one study (Powell, Horne, Pinillos, & Holyoak, 2015) coded participant response quality and reported their results under a "strict" and "liberal" quality requirement, where the former excluded more than the latter. Five studies did not report using any measure of data quality (see Table 1).

Of those that reported it, the median percentage of potential total sample size excluded by these measures was 16%; however, there was notable variability (SD

14.33%). It ranged from 3.51% excluded (Swain, Alexander, & Weinberg, 2008) to 52% (Machery et al., 2015). The latter study tested cross-cultural differences and translated Gettier cases into foreign languages. Considering how many participants were excluded by the comprehension questions, it is worth questioning if something was, for lack of a better term, lost in translation. (They did report that “[t]he translators were bilingual scholars who were native speakers of [their respective languages],” p. 12, Ibid.) In any case, if 16% of participants did not understand the survey prompts, the use of a comprehension question seems necessary given the modest effect size(s) of many of these studies.

There is also little agreement on the best response format in knowledge attribution questions. Of the studies that reported response format, half (n=7) used questions with binary response options. Generally the options provided were "really knows" and "only believes", however that latter was substituted with "does not know" on occasion. Two studies expanded on the binary response options but adding a third category "*S* thinks they know *p*, but does not really know it". Rating scales was used by three studies, both five and seven point, ranging from "knows" to "does not know". The remaining two studies that reported response format used varied knowledge attribution questions, as this a part of their hypotheses.

In general there does not seem to be one type of response format that drives more consistent results. When Cullen replaced "really knows" and "only believes" with "knows" and "does not know," he found knowledge attribution jumped from 28% to 57% (2013). It therefore seems likely that the standard "really knows" and "only believes"

prompts are leading participants to deny knowledge. At the very least, it seems to imply that knowledge is "above" belief in some fashion. While this has been an assumption of philosophers since the original Gettier cases appeared, participants (i.e., laypeople) may or may not subscribe to this epistemic ontology. Adding a third category in-between knows and believes will not prevent a study from leading participants if the phrasing of the knowledge and denial choices are similarly phrased.

While using a rating system might seem to increase sensitivity to small changes in knowledge attribution, they also can make the interpretation of results difficult. For example, how could one meaningfully interpret the difference between a 5 and 6 out of 7 on a scale from "really knows" to "only believes"? When Cullen switched from a dichotomous scale to a seven point rating scale the modal response was the midpoint, however the number of participants attributing knowledge still increased dramatically (2013). This result is not nearly as simple to interpret as a case where only X% of participants attributed knowledge with dichotomous response categories, and may only serve increase complexity for both researcher and participant. Ultimately, there does not seem to be any clear reason to prefer one form of response scoring over the other.

### **Readability Analysis**

Across these 15 studies, 37 different Gettier cases were used (see Appendix B). While the Buick case was used in at least 5 studies, most cases were used in only a single study. The length of the cases varied considerably from nearly 600 words (the Detective case used by Powell, Horne, Pinillos, & Holyoak, 2015), to under 100 words (e.g. the Fake Barn case version 2 used by Swain, Alexander. & Weinberg, 2008). There seems to

be little consistency in this area, and this perhaps is one of the factors contributing to the difficulty replicating findings.

The median Dale-Chall Readability score across all cases was 6.6 (SD = 1.23), suggesting that one would need to read at a 7-8 grade level to comprehend most Gettier cases. However, two cases were found to be significantly different than this median score. The Cancer Conspiracy case was greater than 2 standard deviations above the median with a score of 9.6, suggesting a college-level reading ability requirement. Conversely, the Zebra case version 1 had a readability score of only 1.5, about 2 standard deviations below the mean, suggesting a reading level requirement below the 4th grade. Both cases were used by Weinberg, Nichols, and Stich (2001) and later the replication by Seyedsayamdost (2015). As a result, these studies both had highest range of readability scores (range of 8.1, see Table 2).

Interestingly, Weinberg, Nichols, and Stich (2001) reported a difference on the Cancer Conspiracy case between low and high SES individuals, as measured by education. It is plausible that this difference was due to reading difficulty of the passage; they did not test the Zebra case across SES. Seyedsayamdost (2015) found no effect for either case across SES. Again, neither of these studies used a comprehension question, so their results should be interpreted with caution.

The highest average readability score was in Colaço, Buckwalter, Stich, and Machery's (2014) study. They only tested the Fake-barn case version 1 that has a Dale-Chall Score of 7.5; this suggests a 9-10 grade reading level for the passage. Participants for this study “were recruited in various public places (p.201).” Although over 10 years



old, the National Assessment of Adult Literacy (NAAL) found that over a third of adults had only basic or below basic reading ability for documents, and that reading ability was inversely associated with age. Colaço, Buckwalter, Stich, and Machery (2014) reported the only significant difference in their study was due to age, but considering the reading difficulty of their instrument together with the NAAL's results, that finding seems questionable. Conversely, Starmans and Friedman (2012) had the lowest average readability score (5.68). Since their study endeavored to show a difference between the intuitions of philosophers and laypeople, a lower reading level would be desirable. It is perhaps worth noting that Colaço et al. (2014) did not find an effect of their original hypothesis (that “high defeaters” of knowledge would be less likely “to attribute knowledge”), perhaps their range of readability scores contributed to this.

### **Participant Fatigue**

Approximately half (n=7) of the studies only presented participants with a single thought experiment (See Table 1). For these, the potential for fatigue is circumscribed. However, one study (Nagel, Juan, & Mar, 2013) presented a series of 9 Gettier cases, and another (Adleberg, Thompson, & Nahmias, 2015) presented 7 thought experiments (with a Gettier case being fourth). Another study (Turri, 2013) presented only a single Gettier case to participants, but gave longer versions of it, well over 500 words. It is not clear how presenting multiple cases, or other idiosyncratic changes, might influence participant responses considering the work presented by Cullen (2013).

In a classic study to determine the effect of participant fatigue, Herzog and Bachman found that fatigued participants were more likely to "straight line" responses.

This refers to a propensity to answer with a position bias (such as always selecting the middle option) when fatigued or disengaged. The authors of this paper suggest that participant interest in the topic of the survey might mediate the relationship between fatigue and engagement (1981). While it is certainly possible that participants find the present study more interesting than other survey topics, there does not seem to be hard evidence to this effect (anecdotal report aside).

Aside from the thought experiments, it also possible other survey artifacts could induce participant fatigue. While the prompts for the present studies tended to be short (e.g. "Does *S* really know, or only believe *p*"), a number of studies presented a battery of questions following each thought experiment. Much of the research on participant fatigue, such as the study by Herzog and Bachman (1981), uses only question batteries to study the effect of participant fatigue. Conversely, the studies using Gettier cases tend to alternate between thought experiment and question batteries. It seems plausible that this task-switching could have a greater impact on participant fatigue.

There does not currently seem to be any research on how participant fatigue affects participant response on thought experiments specifically. At the very least, fatigue would likely result in participants giving poorly thought-out responses, and encourage straight-lining, both serious concerns for data quality. In my informal testing of Gettier cases, participants have reported their "brain being mush" after responding to a series of Gettier cases. Again, it is not clear if this affects data quality, but it seems to be a legitimate concern. This concern is not limited to the present studies, as much of experimental philosophy involves testing thought experiments in some form. More work in this area

needs to be done if we truly wish for experimental philosophy using thought experiments without concerns of fatigue.

### **Comprehension questions**

Two-thirds (n=10) studies report using either comprehension questions, or some form of attention check. On average, 16.21% of potential participants were excluded by these measures. For many studies, it seems this exclusion has no effect on the study results. For example, the study by Powell, Horne, Pinillos, and Holyoak employed two different data quality scoring techniques, one strict and one liberal, with the latter excluding more than the former. However, they reported no effect of either on their overall results (2015). John Turri mentioned using comprehension questions in every study he has run (likely hundreds of smaller trials, the vast majority remain unpublished), and exclusion of participants who failed said question(s) effected his results only once (personal communication, September 9, 2016). Conversely, consider the Machery et al. study that excluded 53% of their potential participants - comprehension questions certainly made a difference here! The point is not to debate whether or not comprehension questions *do* affect data quality, only to suggest they *could*. After a study has been run, it would be impossible for researchers to go back and collect data on participant comprehension. If a researcher were to then find a small effect size and/or otherwise messy data, it would be hard to view their results credibly.

In addition, researchers might endeavor to ask more difficult comprehension questions than a simple dichotomous question or two. Asking multiple or open-ended comprehension questions, or at least providing more than two options would likely

increase data quality. For example, one could ask, “what kind of car does Jill drive?” and allow “American, German, Japanese, or Unclear – not enough information provided.”

This would only allow a quarter of those who are randomly guessing to participate. Use of more difficult comprehension questions might also catch more participants who are not paying attention, or truly do not understand the thought experiment, and this might have a greater impact on the findings. Still, even a single, yes/no comprehension question would potentially increase data quality over no comprehension question. Given the difficulty with reproducing results in this field, their use seems warranted.

## Discussion

There is a wide assortment of thought experiments that have been used experimentally under the term “Gettier case.” These vary in length, readability, and purpose to such a degree that it is not altogether clear that they can be considered a single instrument, at least from an experimental standpoint. While at least one thought experiment had a readability score that would imply a college-level reading requirement, eight others also had a score that would imply a mid-to-late high school reading ability (Dale-Chagall Score of 7 or more). This is higher than is suggested by the National Library of Medicine for medical literature, which suggests a 7th or 8th grade reading level (<https://www.nlm.nih.gov/medlineplus/etr.html>). Although the purpose here is obviously different, researchers should not ignore the reading level of the thought experiments they are using, even if their sample consists of college students (who, we hope, are reading at least at a high school level). If, as some experimenters have, researchers intend to give the thought experiments to individuals with a lesser educational background, this issue becomes even more important.

One criticism of readability analysis in general is that it often weights too heavily words with multiple syllables and many letters (Klare & Buck, 1954; Lenzner, 2014). For example, the word *asparagus* would make a passage more difficult than the word

*faux*, even though most would judge the prior, but not the latter, as trivial. The Dale-Chall readability index escapes this to a degree but relying on a list of commonly used words, but it still weights sentence length. A sentence such as “*Man is not without what he has lost,*” is probably more difficult than “*The car could still run, because it had plenty of gasoline and the engine worked fine.*” Gettier cases would likely fall into the camp of passages which are fairly easy to read, but less easy to actually understand. Although the computable readability may not be high for every case, one should be cognizant that these thought experiments are probably difficult for many to understand. Indeed, this is what makes them interesting in the first place.

Further evidence that some participants might be having difficulty understanding the thought experiments comes from the finding that, on average, 15% are excluded by the use of comprehension questions. And while we hope that those who are not excluded do actually understand the critical concepts in play, it is always possible that they could only continue through sheer luck. Or perhaps the comprehension questions are too easy, such that one could answer them correctly without actually understanding the threats to the agent's knowledge. Researchers should be continually looking to improve these comprehension questions, and any study that does not employ them will certainly have questionable data quality. Again, while their use may not be necessary for every study, future researchers should employ them nonetheless. Their use should cause minimal strain on participants, and give researchers an option to remove participants' data that might skew their results.

Although many studies were found to only use a couple of thought experiments at once, some employed many. In addition, some studies used very lengthy passages and/or filler studies. All of this can contribute to participant fatigue and issues of reading difficulty will only compound this. However, one of the weaknesses of the present study is that much of the discussion surrounding participant fatigue is largely speculative. At present, little is known exactly if this affects data quality for epistemological studies, or precisely when participants start to become fatigued. This is an area for further research.

Although not discussed in great detail in the present study, Cullen (2013) provided an interesting set of findings regarding the survey prompts. For example, changing “Really knows” and “Only believes” to “Knows” and “Does not know.” This can apparently have fairly drastic effects on participant response. He also asks, quite plausibly, if the standard prompts of “Really knows” and “Only believes” are leading. Simply because philosophers have decided to place a higher weight on knowledge over belief, we should not assume that laypeople will do the same. This, too, is an area which researchers should investigate further, and when designing studies, should question if their prompts are most appropriate. On the other hand, drastically changing prompts between studies will make comparison difficult, if not impossible. Researchers should carefully weigh their options here and must decide which is most appropriate given their purpose.

The studies were found to use a myriad of thought experiments that have divergent purposes. It might be preferable to use particular cases more consistently to facilitate comparison between studies. However, many studies tweaked the cases

purposefully for idiosyncratic reasons and using the same cases might not have made sense for them. In any case, there are a few clear recommendations to improve data quality, and hopefully reduce the number of unreplicable findings. Studies should: use comprehension questions, try to use cases where the reading level is appropriate, counterbalance when presenting multiple cases, and present as few cases as possible. Again, these recommendations will not be appropriate for every study, and researchers will, of course, make changes to fit their particular needs. However, as a general set of principles this should suffice. More work needs to be done, to increase the specificity of these principles, particularly in the area of participant fatigue and survey prompts. Finally, little (if any) work has been done regarding the reliability of Gettier cases. Ostensibly, researchers hope to find responses that are representative of enduring intuitions, not mere passings of thought. This too is an area that could use more study.

Based on this analysis, there are five general recommendations to future researchers working with Gettier cases. The first is to use comprehension questions. On average, about 15% of potential respondents are excluded by their use. This is not an insignificant number, and could seriously skew responses in aggregate. Second, researchers should counterbalance their studies when working with multiple thought experiments. Cullen's (2013) work provides rather strong evidence to suggest that they are sensitive to order effects. Third, researchers should check the readability of their thought experiments, especially when working with new ones. Ideally, the range of readability across all thought experiments used should be low, and differences acknowledged. Fourth, researchers should attempt to minimize participant fatigue.



Philosophical thought experiments can be mentally taxing, and it is unclear how this affects data quality.

The final recommendation is somewhat more speculative than the rest: researchers should ask an open-ended, qualitative question asking participants to justify their results or explain their reasoning. There are two main reasons for this. First, participants might pass the comprehension questions but still might not understand the critical components of the thought experiment. This determination will be a somewhat subjective call on the part of the researcher(s) and probably should not be exercised except in rare cases. It is, of course, questionable whether this would affect results, but again the researcher(s) might prefer to at least have the option.

The second justification, however, is the more pressing concern. Qualitative analysis of participants' justifications might lead philosophers to consider the cases differently. Many participants will be approaching these thought experiments from an entirely different perspective than that of professional philosophers. Consider the discussion in Simon Cullen's third experiment using the Truetemp case where participants typically cited internalist arguments as a motivation to deny knowledge (the standard philosophical response). One participant "cited [the agent's] *ignorance* in support of the opposite, distinctly externalist conclusion!" (p. 290, 2010, emphasis in original). The idea that only professional philosophers can have valid insights on the nature of these cases approaches a dangerous form of academic elitism. Ostensibly, we are interested in the intuitions of our participants in these studies, and we should not limit

these intuitions to a few options on a survey. These cases are far too complex, and the intuitions of laypeople can be as well.

Experimentally testing Gettier cases could continue to prove valuable to both philosophy and psychology. However, the studies should be carefully planned and interpreted. The number of unreplicable findings in this area is worrisome, empirically speaking. Robust findings will require further work on the use of Gettier cases methodologically before we hope to answer our theoretical questions.

## References

- Adeberg, T., Thompson, M., & Nahmias, E. (2014). Do men and women have different philosophical intuitions? Further data. *Philosophical Psychology*, 28(5), 615-641.  
doi:10.1080/09515089.2013.878834
- The analysis of knowledge (Stanford encyclopedia of philosophy). (2001, February 6).  
Retrieved from <http://stanford.library.sydney.edu.au/entries/knowledge-analysis/>
- Beebe, J., R., & Undercoffer, R., J. (in press). Individual and cross-cultural differences in semantic intuitions: New experimental findings. *Journal of Cognition and Culture*. Retrieved from:  
<http://www.acsu.buffalo.edu/~jbbeebe2/Beebe%20Undercoffer%20Individual%20and%20Crosscultural%20Differences%20in%20Semantic%20Intuitions.pdf>
- Buckwalter, W. (2013). Gettier made ESEE. *Philosophical Psychology*, 27(3), 368-383.  
doi:10.1080/09515089.2012.730965
- Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.
- Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic Intuitions in Fake-Barn thought experiments. *Episteme*, 11(02), 199-212.  
doi:10.1017/epi.2014.7

- Cullen, S. (2013). Survey-driven romanticism. *Review of Philosophy and Psychology, 1*(2), 275-296. doi:10.1007/s13164-009-0016-1
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis, 23*(6), 121-123. doi:10.1093/analys/23.6.121
- Goldman, A. I. (1976). Discrimination and perceptual knowledge. *Journal of Philosophy, 73*(20), 771-791. doi:10.2307/2025679
- Herzog, A. R., & Bachman, J. G. (1981). Effects of questionnaire length on response quality. *Public Opinion Quarterly, 45*(4), 549. doi:10.1086/268687
- How to write easy-to-read health materials: MedlinePlus. (2016, March 23). Retrieved July 1, 2016, from <https://www.nlm.nih.gov/medlineplus/etr.html>
- Kim, M., & Yuan, Y. (2014). No cross-cultural differences in the Gettier car case intuition: a replication study of Weinberg et al. 2001. *Episteme, 12*(3), 355-361.
- Klare, G. R., & Buck, B. (1954). Formula pitfalls. In *Know your reader: The scientific approach to readability* (pp. 136-151). New York, NY: Hermitage House.
- Lenzner, T. (2013). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research, 43*(4), 677-698. doi:10.1177/0049124113513436
- Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., ... Hashimoto, T. (2015). Gettier across cultures. *Noûs*. doi:10.1111/nous.12110
- Nagel, J., Juan, V. S., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition, 129*(3), 652-661. doi:10.1016/j.cognition.2013.02.008

- National assessment of adult literacy (NAAL). (n.d.). Retrieved June 26, 2016, from [https://nces.ed.gov/naal/kf\\_demographics.asp](https://nces.ed.gov/naal/kf_demographics.asp)
- Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological review*, *108*(2), 291-310.
- Powell, D., Horne, Z., Pinillos, N. Á., & Holyoak, K. J. (2015). A Bayesian framework for knowledge attribution: Evidence from semantic integration. *Cognition*, *139*, 92-104. doi:10.1016/j.cognition.2015.03.002
- Seyedsayamdost, H. (2015). On normativity and epistemic intuitions: Failure of replication. *Episteme*, *12*(1), 95-116.
- Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition*, *124*(3), 272-283. doi:10.1016/j.cognition.2012.05.017
- Stich, S. (2012). Do different groups have different epistemic intuitions? A reply to Jennifer Nagel 1. *Philos Phenomenol Res*, *87*(1), 151-178. doi:10.1111/j.1933-1592.2012.00590.x
- Swain, S., Alexander, J., & Weinberg, J. M. (2008). The instability of philosophical intuitions: Running hot and cold on truetemp. *Philosophy and Phenomenological Research*, *76*(1), 138-155. doi:10.1111/j.1933-1592.2007.00118.x
- Turri, J. (2016). Knowledge judgments in “Gettier” cases. *A Companion to Experimental Philosophy*, 335-348. doi:10.1002/9781118661666.ch23

- Turri, J. (2013). A conspicuous art: Putting Gettier to the test. *Philosophers' Imprint*, 13(10), 1-16. Retrieved from <http://quod.lib.umich.edu/cgi/p/pod/dod-idx/conspicuous-art-putting-gettier-to-the-test.pdf?c=phimp;idno=3521354.0013.010>
- Turri, J., Buckwalter, W., & Blouw, P. (2014). Knowledge and luck. *Psychonomic Bulletin & Review*, 22(2), 378-390. doi:10.3758/s13423-014-0683-5
- Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical topics*, 49(1&2), 429-460.
- Ziółkowski, A. (2015). Folk intuitions and the No-Luck thesis. *Episteme*, 1, 1-16. doi:10.1017/epi.2015.49

## Appendix A

### Paradigmatic Gettier Cases

Car case (or Buick case):

“Bob has a friend, Jill, who has driven a Buick for many years. Bob therefore thinks that Jill drives an American car. He is not aware, however, that her Buick has recently been stolen, and he is also not aware that Jill has replaced it with a Pontiac, which is a different kind of American car. Does Bob really know that Jill drives an American car, or does he only believe it?”

(Weinberg, Nichols, & Stich 2001, p. 449)

Fake barn case:

“Henry is driving in the countryside, looking at objects in fields. He sees what looks exactly like a barn. Accordingly, he thinks that he is seeing a barn. Now, that is indeed what he is doing. But what he does not realize is that the neighborhood contains many fake barns — mere barn facades that look like real barns when viewed from the road. And if he had been looking at one of them, he would have been deceived into believing that he was seeing a barn. Luckily, he

was not doing this. Consequently, his belief is justified and true. But is it knowledge?"

("Fake barn" case; Goldman, 1976)



## Appendix B

### Alphabetical Listing of Gettier Cases

The italicized prompts are standardized in the form *Does S really know p, or only believe it?* They were included in the readability analysis, but not necessarily presented in the study.

#### Baby Deer Case:

Brad is driving up to his cousin's cottage north of the city with some friends of his. It is late Friday afternoon, and it's sunny and clear out. At one bend in the road Brad points at something on the side of the road. "Hey, look it's a baby deer!" Everyone in the car sees what Brad does, and agrees that there's a very cute baby deer standing very still in the fenced field by the side of the road. Actually what they were looking at wasn't a deer at all but a very realistic statue put up for decoration by the property owner. They slow down to take a good look, and keep on driving. By chance there was also a real deer in the field by the road, but they didn't see it - it was hidden just out of their sight behind some trees.

*Do they know there is a baby deer in the field, or only believe it?*

Dale-Chall Readability score: 5.3

Presented by: Nagel, Juan, & Mar (2013)

Bookstore Case:

Jane owns a small bookstore. One morning, while a few customers are browsing through the books, Jane puts a \$20 bill into the empty cash register. She locks is up and goes into the back room to brew some coffee. While Jane is in the back room, her employee Bill arrives. He looks in the register and notices that the \$20 is a little worn. He takes it out of the cash register and replaces it with a new \$20 bill and leaves to run

some errands. Jane has only been in the back room for a few minutes and did not hear anything.

*Does Jane know there is a \$20 in the register, or only believe it?*

Dale-Chall Readability score: 5.3

Presented by: Starmans & Friedman (2012)

Broken Clock Case Version 1:

Mary works as a clerk in an office, she is clear-headed and has excellent eye-sight. Mary knows that she set the clock above her desk accurately and that it has been completely reliable for many years. At 3:00pm Mary looks up at the clock and sees that it reads “3:00pm”, and indeed, it is 3:00pm. However, unknown to Mary, the clock stopped working

exactly 24 hours ago.

*Does Mary really know that the time is 3:00pm, or does she only believe it?*

Dale-Chall Readability score: 5.4

Presented by: Cullen (2013)

Broken Clock Case Version 2:

Wanda is out for a weekend afternoon walk. As she passes near the train station, she wonders what time it is. She glances up at the clock on the train station wall and sees that it says 4:15 pm. What she doesn't realize is that this clock is broken and has been showing 4:15 pm for the last two days. But by sheer coincidence, it is in fact 4:15 pm just at the moment when she glances at the clock.

*Does Wanda really know it is 4:15 pm, or only believe it?*

Dale-Chall Readability score: 5.4

Presented by: Nagel, Juan, & Mar (2013)

Buick Case:

Bob has a friend, Jill, who has driven a Buick for many years. Bob therefore thinks that Jill drives an American car. He is not aware, however, that her Buick has recently been stolen, and he is also not aware that Jill has replaced it with a Pontiac, which is a different kind of American car.

*Does Bob really know that Jill drives an American car, or believe it?*

Dale-Chall Readability score: 6.1

Presented by: Weinberg, Nichols, & Stich (2001); Seyedsayamdost (2015); Kim & Yuan (2014); Nagel, Juan, & Mar (2013); Cullen (2013)

Note: In some studies, Bob drives a Japanese, rather than American car with the makes changed accordingly. This was done under the assumption that East Asian participants may be more familiar with Japanese makes. However the difference in quantifiable readability is negligible.

Cancer Conspiracy Case:

It's clear that smoking cigarettes increases the likelihood of getting cancer. However, there is now a great deal of evidence that just using nicotine by itself without smoking (for instance, by taking a nicotine pill) does not increase the likelihood of getting cancer. Jim knows about this evidence and as a result, he believes that using nicotine does not increase the likelihood of getting cancer. It is possible that the tobacco companies dishonestly made up and publicized this evidence that using nicotine does not increase the likelihood of cancer, and that the evidence is really false and misleading. Now, the tobacco companies did not actually make up this evidence, but Jim is not aware of this fact.

*Does Jim really know that using nicotine doesn't increase the likelihood of getting cancer, or does he only believe it?*

Dale-Chall Readability score: 9.6

Presented by: Weinberg, Nichols, & Stich (2001); Seyedsayamdost (2015)

Coin Case Version 1:

Corey has been collecting coins in his piggy bank for years. One day he is about to put a quarter in his piggy bank, and notices that it looks pretty old. Though he's never paid attention to dates before, he reads the date and sees that it's from 1936. However, he doesn't realize that the date has partially rubbed off and it is really from 1938. There is already a quarter dated 1936 buried deep in his piggy bank, but Corey isn't aware of this. He deposits the quarter and goes to take a nap. Corey's roommate Scott comes home, and needs some change for the bus. He shakes the piggy bank and the quarter Corey just put in falls out. Scott takes it and leaves. Corey wakes up after a 10-minute nap, and doesn't realize that Scott was there.

*Does Corey really know there is a quarter dated 1936 in his piggy bank, or only believe it?*

Dale-Chall Readability score 6.4

Presented by: Starmans & Friedman (2012)

Coin Case Version 2:

Robert recently made a purchase for a rare 1804 US silver dollar. He keeps the coin on display over the fireplace in his library. This evening Robert is having his neighbors over for dinner. He puts the coin in its display over the fireplace, shuts the library doors behind him, and hurries to greet his guests, who just arrived. He greets them and says, "Guess what? There is an 1804 US silver dollar in my library." When Robert shut the library doors, a coin thief silently entered through the library window, stole Robert's 1804 US silver dollar, and quickly escaped. Robert had only been out of the library for a few seconds and did not hear anything. The coin was already gone by the time Robert greets his guests and tells them, "There is an 1804 US silver dollar in my library." Robert's house is a very old mansion. Back in the early 1800s, when the house was originally built, one of the carpenters accidentally, and without noticing, dropped an 1804

US silver dollar into the mortar mix used to make the fireplace. This lost silver dollar is still in the fireplace in the library. But no one has seen it for hundreds of years, and no one will ever see it again. It will remain hidden in Robert's library.

*Does Robert really know there is a 1804 US silver dollar in his library, or only believe it?*

Dale-Chall Readability score: 5.7

Presented by: Turri (2013)

#### Conflict Diamond Case:

Emma was shopping for diamonds. Shopping was always exciting, but she was worried about conflict diamonds—illegal diamonds from war-torn regions of Africa that are mined with slave labor and fund violent militias and warlords.

Emma had a friend, Jim, who worked at Amnesty International, a humanitarian organization that monitors conflict diamonds. Emma called Jim to ask how she might avoid conflict diamonds. He explained that legitimate diamond mines assign serial numbers to all their diamonds that certify the diamonds as conflict-free. If she found a diamond she liked, Jim told Emma to call him so he could look up the diamond's serial number in Amnesty International's database. Then he could send her a certificate of authenticity so she'd have proof the diamond was conflict-free.

The next day, Emma went to a jewelry shop and found a beautiful diamond necklace that she really liked. She checked the paper tag for the diamond's serial number and wrote it down.

Emma called her friend and asked him to look up the serial number for the diamond. Jim began to look the diamond up in Amnesty International's database but he made a typo while he was typing in the serial number. He accidentally

looked up the wrong diamond, but found that it was listed as a conflict-free diamond. Jim sent Emma the certificate of authenticity for this other diamond. As luck would have it, Emma's diamond actually was conflict-free. She was thrilled when she received the certificate that Jim had sent. Emma didn't feel one bit guilty about purchasing the diamond, since she thought it was conflict-free.

Emma wanted to take her friend Jim out for lunch to thank him. Just then, she got a call from her law firm. She had just been made a junior partner, and that meant she was almost always on call. The call was important, it sounded like it could mean a major break in her case. That was good news, but it meant she would be very busy for the next few days. She told her paralegal assistant to get the case files ready for her at the office. It looked like lunch with her friend would have to wait.

*Does Emma really know her diamond is conflict-free, or only believe it?*

Dale-Chall Readability score: 6.8

Presented by: Powell, Horne, Pinillos, & Holyoak (2015)

Detective Case:

Gary Hawkins was a counselor who treated troubled youths with long histories of abuse. He was having an especially hard time getting through to two of his clients, a pair of fourteen year-olds named Will and Beth, who both seemed to dislike him. Most of Gary's clients grew up poor and were at-risk youths.

One morning, Gary was out for a jog in Millennium Park on the east side of Chicago. Gary's path ran under Columbus Drive, and when he entered the unlit tunnel his eyes were unadjusted to the dark. Suddenly, Gary felt a terrible pain at the back of his head and he fell to the ground. He hadn't seen the attacker waiting in the tunnel with a weapon in their hand. The attacker continued to hit Gary with

the weapon, bruising his ribs and arms. Then the attacker ran off, and Gary laid in the tunnel, dazed.

Another jogger discovered Gary about a half an hour later and called the police. Detective Jack Dempsey was assigned to the case. Dempsey was a veteran detective who loved police-work, so he hurried to the hospital to interview Gary as soon as his doctors would allow it. Unfortunately, Gary was useless as a witness. He hadn't seen the attack coming, and the blow to the head had left his memory hazy. Next, Dempsey started to question Gary's clients, and Will really rubbed him the wrong way. Dempsey was immediately suspicious of him.

Dempsey wasn't the only one who disliked Will. Beth and Will were dating, and she suspected he was going to leave her. She wanted a way to get even with Will, and Will had told her a couple weeks before that he was planning to attack Gary in Millennium park.

Dempsey started his investigation and found several pieces of evidence that pointed to Will. First, another officer found Will's baseball bat near the scene of the crime. Then, Dempsey got a warrant and searched Will's phone, where he found texts bragging about beating Gary up.

Actually, Beth wanted to get payback for Will leaving her. She wanted to make sure Will was caught for his crime. Will was careful to cover his tracks after he attacked Gary. He attacked Gary with a pipe that he disposed of in the lake, and he never sent any texts about the incident. Beth was getting payback for Will leaving her by planting evidence! She sent the texts from Will's phone, and planted Will's baseball bat at the crime scene.

Beth's deception ensured that Dempsey would have accused Will of the crime even if he had been completely innocent. Dempsey had the right suspect, but only through luck.

After finishing his investigation, Dempsey wrote up his report for the district attorney based upon the evidence he had collected, including Beth's



testimony. He worked on his other cases until Will's case went to trial. Whatever the ultimate verdict would be, Dempsey thought Will was guilty.

Dempsey tried not to worry about work and just look forward to the weekend. His daughter was visiting colleges, and they were flying to New York together to visit NYU. Dempsey had never visited New York before, and he really needed a vacation. It would be a good chance for a break, although he kept warning his daughter that Chicago's pizza was vastly superior.

*Does Dempsey really know that Will was guilty or only believe it?*

Dale-Chall Readability score: 7.3

Presented by: Powell, Horne, Pinillos, & Holyoak (2015)

#### Diamond Case Version 1:

Emma is shopping for jewelry. She goes into a nice-looking store, and selects a diamond necklace from a tray marked "Diamond Earrings and Pendants". "What a lovely diamond!" she says as she tries it on. Emma could not tell the difference between a real diamond and a cubic zirconium fake just by looking or touching. In fact, this particular store has a very dishonest employee who has been stealing real diamonds and replacing them with fakes; in the tray Emma chose almost all of the pendants had cubic zirconium stones rather than diamonds (but the one she chose happened to be real).

*Does Emma really know she has an actual diamond, or only believe it?*

Dale-Chall Readability score: 6.6

Presented by: Nagel, Juan, & Mar, 2013, Machery et al. (2015), Powell, Horne, Pinillos, & Holyoak (2015)

Note: in the Machery et al. (2015) study, the case is slightly different (and not presented as a "true" Gettier case), but the Dale-Chall Readability score only changes marginally (a decrease of 0.30).

#### Diamond Case Version 2:

Emma is admiring jewelry in a fancy department store. She is particularly fascinated by the stones in the diamond display. After discussing it with the sales associate, she selects a stone, pays for it, and puts it in her pocket. After browsing for another minute, Emma leaves the store. Unfortunately, the stone Emma bought is a fake. It's not a real diamond. It's a worthless cubic zirconium. While Emma was browsing, she didn't notice that a stealthy pickpocket tried to reach into her pocket and steal the fake stone she just bought. The pickpocket succeeded. Another thing happened while Emma was browsing. Her disguised ex-husband was also in the store. In order to frame Emma for robbery, he stealthily slipped a stolen diamond into Emma's pocket. No one, not even Emma, noticed that this happened.

*Does Emma really know she has a diamond in her pocket, or only believe it?*

Dale-Chall Readability score 6.1

Presented by: Turri, Buckwalter, & Blouw (2015)

#### Emmy Case:

Peter and his girlfriend are big fans of Naomi Watts. Tonight, Peter is watching the Emmy Awards on television, and when Naomi comes on stage he calls out to his girlfriend, who is in the next room, "Come on, Naomi Watts is on stage! Indeed, Peter is right it is Naomi Watts on the stage. However, if it had been Scarlett Johansson, who looks very similar to Naomi, Peter would have mistaken her for Naomi Watts.

*Does Peter really know that Naomi Watts is on stage or does he only believe it?*

Dale-Chall Readability score: 8.0

Presented by: Cullen (2013)

Fake Barn Case Version 1:

Gerald is driving through the countryside with his young son Andrew. Along the way he sees numerous objects and points them out to his son. ‘That’s a cow, Andrew,’ Gerald says, ‘and that over there is a house where farmers live.’ Gerald has no doubt about what the objects are. What Gerald and Andrew do not realize is the area they are driving through was recently hit by a very serious tornado. This tornado did not harm any of the animals, but did destroy most buildings. In an effort to maintain the rural area’s tourist industry, local townspeople built house façades in the place of destroyed houses. These façades look exactly like real houses from the road, but are only for looks and cannot be used as actual housing.

Though he has only recently entered the tornado-ravaged area, Gerald has already encountered a large number of house façades. However, when he tells Andrew ‘That’s a house,’ the object he sees and points at is a real house that has survived the tornado.

*Does Gerald really know he is looking at a real house, or only believe it?*

Dale-Chall Readability score: 7.5

Presented by: Colaço, Buckwalter, Stich, & Machery (2014); Ziółkowski (2015)

Note: This version was presented to the "high-defeaters" experimental group for the Colaço, Buckwalter, Stich, & Machery (2014) study; the "low-defeaters" group saw a slightly different version with a marginally higher Dale-Chall

Readability score of 7.7; in the latter condition, Gerald has yet to encounter the house façades before his comment.

#### Fake Barn Case Version 2:

Suzy looks out the window of her car and sees a barn near the road, and so she comes to believe that there's a barn near the road. However, Suzy doesn't realize that the countryside she is driving through is currently being used as the set of a film, and that the set designers have constructed many fake barn facades in this area that look as though they are real barns. In fact, Suzy is looking at the only real barn in the area.

*Does Suzy really know she is looking at a real barn, or only believe it?*

Dale-Chall Readability score: 6.0

Presented by: Swain, Alexander. & Weinberg (2008)

#### Furniture Store Case:

Albert is in a furniture store with his wife. He is looking at a bright red table in a display. He believes the table is just the shade of red he was looking for. However, a white table under red lighting would look *exactly* the same to him, and Albert has not checked whether the lighting in this store is currently normal. In fact, this showroom was set up by a very creative lighting consultant who has put different brightly-colored spotlights over almost all the furniture on display. Most things in the store are lit so that they appear to be different colors, but not the one table that Albert is now looking at. The colored spotlights are set up so that shoppers don't notice them at first. Albert asks his wife "Do you like this red table?"

*Does Albert know the table is red, or only believe it?*

Dale-Chall Readability score 5.9

Presented by: Nagel, Juan, & Mar (2013); Machery et al. (2015)

Note: In the Machery et al. (2015) study, this is not presented as a Gettier case, but rather a case of "clear knowledge" and is accordingly edited such that there is no threat to Albert's knowledge from red lighting. The Dale-Chall Readability score is higher in this form at 10.26

#### Hospital Case:

Paul Jones was worried because it was 10 pm and his wife Mary was not home from work yet. Usually she is home by 6 pm. He tried her cell phone but just kept getting her voicemail. Starting to worry that something might have happened to her, he decided to call some local hospitals to ask whether any patient by the name of "Mary Jones" had been admitted that evening. At the University Hospital, the person who answered his call confirmed that someone by that name had been admitted with major but not life-threatening injuries following a car crash. Paul grabbed his coat and rushed out to drive to University hospital. As it turned out, the patient at University Hospital was not Paul's wife, but another woman with the same name. In fact, Paul's wife had a heart attack as she was leaving work, and was at that moment receiving treatment in Metropolitan Hospital, a few miles away.

*Does Paul really know that his wife is in the hospital, or only believe it?*

Dale-Chall Readability score 6.7

Presented by: Nagel, Juan, & Mar (2013); Machery et al. (2015)

Note: The versions presented in the Nagel, Juan, & Mar's (2013) study are slightly different than in Machery et al.'s (2015) study, but the difference is negligible from a readability standpoint.

### ID Card Case:

Amanda has just arrived at Atlantic University as a spring term transfer student. She needs to get a library card, and goes to the information desk at the library to ask where to get a card. There is no one staffing the desk just then, so she asks a passing student instead. The student tells her that campus ID cards, which also work as library cards, are issues at a booth on the ground floor of the student activity center. Amanda thanks the student and heads over to the student activity center. As it turns out, this student is confused - he got his ID card at a booth at that location during Orientation Week last fall, but that booth was taken down at the end of the week. At other times in the year students are supposed to get their ID cards from an office on the second floor of the library. However it just so happens the computer network in the library is down today, and so just for today the ID card staff have set up their temporary booth over in the student activity center.

*Does Amanda really know she can get an ID card in the student activity center, or only believe it?*

Dale-Chall Readability score: 6.7

Presented by: Nagel, Juan, & Mar (2013)

### Las Vegas (Vacation Trip) Case:

Luke works in an office in New York with two other people, Victor and Monica. All winter Victor has been describing his plans to go to Las Vegas on his vacation, even showing Luke the website of the hotel where he has reservations. When Victor is away on vacation, Luke receives a very nice email from Victor together with photos of Victor posing in front of Las Vegas landmarks. When he gets back to work, Victor talks a lot to Luke about how much fun he had

vacationing in Las Vegas. However, Victor didn't really go on the trip; he has just been pretending. His tickets and reservations were cancelled because his credit card was maxed out, and he secretly stayed home in New York, very skillfully faking the photos he sent Luke. Meanwhile, Monica just spent a weekend vacationing in Las Vegas, but kept this a secret from all her co-workers.

*Does Luke know one of his co-workers vacationed in Vegas, or only believe it?*

Dale-Chall Readability score: 7.4

Presented by: Nagel, Juan, & Mar (2013); Machery et al. (2015)

#### Mall Robbery Case:

Sarah is out with some friends, shopping for clothing at a large mall. As she is standing at the entrance of her favorite store, Sarah sees someone being chased by several security guards. "Oh my God!" she exclaims. "I recognize him. It's Walter, the guy I worked with as a lifeguard last summer." The guards catch up and put their suspect into handcuffs right in front of Sarah and her friends. In fact the person Sarah sees getting caught is not Walter but twin brother Tom. Sarah didn't know that Walter had a twin. As it happens, Walter was also involved in the robbery, but ran the other way, and has been caught by security guards at the other end of the mall.

*Does Sarah really know that Walter has been caught by security guards, or only believe it?*

Dale-Chall Readability score: 6.7

Presented by: Nagel, Juan, & Mar (2013)

#### Mayor Case:

The mayor of a small town is trying to decide whether or not to sign a new contract with a local corporation. The math is all very complex, but all his economic strategists think that there's a relatively good chance that one outcome is that it will cut jobs for workers in the community. The mayor says, "all I really care about is campaign contributions, not people's jobs, and I am sure to get millions from the corporation if I agree." So, he decides to sign their contract. The corporation, however, didn't take any chances. They secretly switched the contract with a totally different one right before the mayor signed it. By changing all the fine print, in some cases the opposite of what the mayor thought he was signing, the corporation could be sure it got what it wanted. Sure enough, shortly after the mayor signed the contract, a number of members of the community lost jobs, and the mayor received a huge donation to his reelection campaign.

*Did the mayor really know the contract would result in lost jobs, or only believe it?*

Dale-Chall Readability score: 6.6

Presented by: Buckwalter (2014)

Note: To test the "epistemic side effect", wherein individuals conceptualize epistemically similar but morally disparate scenarios differently, some respondents saw "cut" and "lost" replaced with "create" and "got". The effect on readability is negligible.

#### Pen Burglar Case Version 1:

Katie is in her locked apartment writing a letter. She puts the letter and her blue Bic pen down on her coffee table. Then she goes into the bathroom to take a shower. As Katie's shower begins, two burglars silently break into the apartment. One burglar takes Katie's blue Bic pen from the table. But the other burglar absentmindedly leaves his own identical blue Bic pen on the coffee table. Then the burglars leave. Katie is still in the shower, and did not hear anything.



*Does Katie really know there is a blue Bic pen on the table, or only believe it?*

Dale-Chall Readability score: 5.4

Presented by: Starmans & Friedman (2012); Turri (2013)

#### Pen Burglar Case Version 2:

Katie is in the living room of her locked apartment writing a letter with a blue Bic pen. She puts the letter and the blue Bic pen down on her coffee table. Then she goes into the bathroom to take a shower. It takes her fifteen minutes to finish. Just after Katie started her shower, two burglars, a master and his apprentice, broke into her apartment. As they made their way around the apartment, the master burglar stole Katie's blue Bic pen from the coffee table. After five minutes, the burglars left, well before Katie finished her shower. Katie did not hear anything. Right before the burglars left Katie's apartment, the apprentice burglar started feeling a bit dizzy, so he sat down on the couch for a moment to recover. When the apprentice burglar sat down, he absentmindedly set his own blue Bic pen on the coffee table, and forgot it there. This was five minutes before Katie finished her shower.

*Does Katie really know there is a blue Bic pen on the table, or only believe it?*

Dale-Chall Readability score: 5.9

Presented by: Turri (2013)

#### Poison Gas Case:

Karen is a distinguished professor of chemistry. This morning, she read an article in a leading scientific journal that mixing two common floor disinfectants, Cleano Plus and Washaway, will create a poisonous gas that is deadly to humans.

In fact, the article is correct: mixing the two products does create a poisonous gas. At noon, Karen sees a janitor mixing Cleano Plus and Washaway and yells to him, “Get away! Mixing those two products creates a poisonous gas!”

*Does Karen really know that mixing the chemicals will create a poison gas, or only believe it?*

Dale-Chall Readability score: 8.0

Presented by: Swain, Alexander. & Weinberg (2008)

#### Pump Case:

Sam’s job is to pump water into the cistern, which then supplies the water to the farms owned by several families in the community. One day, as Sam operates the pump, he hears a broadcast on the radio. The radio report says that local officials suspect a new chemical from a nearby factory, chemical X, may have found its way into the local reservoir, and that there is a chance it will be very beneficial to all the local townspeople’s crops. Sam thinks to himself, “I don’t care about their crops; I just want to earn my pay,” and continues pumping the water. Sure enough, the crops started thriving. It turned out that the local officials were completely wrong about the chemical in the water. After analyzing the water, they found no trace of chemical X. Scientific reports later confirmed that the crops were all thriving because of a fungus that had been secretly growing inside Sam’s pump.

*Did Sam really know that pumping his water would cause the crops to thrive, or only believe it?*

Dale-Chall Readability score: 6.8

Presented by: Buckwalter (2014)

Note: To test the "Epistemic side effect", where individuals conceptualize epistemically similar but morally disparate scenarios differently, some respondents saw "beneficial" and "thriving" replaced with "poisonous" and "dying". The effect on readability is negligible.

#### Special Feeling Case:

Dave likes to play a game with flipping a coin. He sometimes gets a "special feeling" that the next flip will come out heads. When he gets this "special feeling," he is right about half the time, and wrong about half the time. Just before the next flip, Dave gets that "special feeling," and the feeling leads him to believe that the coin will land heads. He flips the coin, and it does land heads.

*Did Dave really know that the coin was going to land heads, or did he only believe it?*

Dale-Chall Readability score: 5.4

Presented by: Weinberg, Nichols, & Stich (2001); Swain, Alexander. & Weinberg (2008); Cullen (2013)

#### Squirrel Case:

Darrel is an ecologist collecting data on red speckled ground squirrels in Canyon Falls national park. The park is divided into ten zones and today Darrel is working in Zone 3. While scanning the river valley with his binoculars, Darrel sees a bushy-tailed creature with distinctive red markings on its chest and belly. The red speckled ground squirrel is the only native species with such markings. Soon Darrel packs up his gear and hikes back to base camp. When Darrel returns to camp, his colleague says, "A reporter is going to do a story on local wildlife and she needs some video footage of a red speckled ground squirrel. Are there

any in Zone 3?" Darrel replies, "Yes, there is at least one red speckled ground squirrel in Zone 3."

The animal Darrel saw in the river valley is a female red speckled ground squirrel, recently photographed by campers. But while Darrel was hiking back to camp, the female he saw migrated out of Zone 3 and never returned. However, a different female, which Darrel never saw, migrated into Zone 3 and made her nest in that same river valley where Darrel was. So there is a red speckled ground squirrel in Zone 3 after all.

*Does Darrel really know that there is a red speckled squirrel in Zone 3, or only believe it?*

Dale-Chall Readability score: 6.7

Presented by: Turri, Buckwalter, & Blouw (2015)

Stopped-Clock case:

John is a watchmaker and has just finished another one of his famous handmade Fregeant-watches. To set the watch, John looks at a clock in his living room and sets the watch after it. However, unbeknownst to John, the clock in his living room had stopped exactly 12 hours before John looked at it. Even though the clock is not running anymore, it is indicating a correct time at this particular moment.

Thus, from now on, the new made watch indicates the correct time. These kinds of watches need not be set again for a long time. John then puts the watch into his display box containing 8 other watches of the same series. The other watches have been set earlier, after the same clock but when it was still running properly.

One hour later, Simon enters the store, wanting to buy a handmade watch from the Fregeant series. Simon picks the one John had just put into the display. To Simon, they all look the same. Simon doesn't set the watch because John tells him that it has just been set. Before Simon leaves the store, John shows him the workshop in which he makes the watches and explains why Fregeant-watches are so precise and reliable.

Next morning, Simon consults the watch for the first time after putting it on his wrist. Given the great reputation of Fregeant watches, Simon has no doubts and comes to truly believe that it is 7.30 am.

*Does Simon really know that it is 7:30 am, or only believe it?*

Dale-Chall Readability score: 6.7

Presented by: Ziółkowski (2015)

Thermometer case:

Stephen is a physician. A patient, Josh, enters his consulting room. Josh feels ill and looks for a correct diagnosis and appropriate treatment. He complains about having a cough, runny nose, and a rash on his skin.

Stephen examines the symptoms of Josh's disease in detail. He discovers that most of the patient's body is covered with small red spots and that his eyes are red as well. These symptoms, together with dry cough and runny nose make Stephen suspect that Josh is suffering from measles. Stephen wants to check whether Josh also has fever.

Stephen takes a box full of thermometers out of his medical cabinet. He picks one thermometer from the box and takes Josh's temperature. The thermometer indicates 98.6 degrees, leading Stephen to believe that the patient's temperature is normal, which is true. However, unbeknownst to Stephen, the

thermometer he used is the only reliable thermometer in the box. All other thermometers in the box are defective and they would read 98.6 even if the patient had a fever. Still, the thermometer that Stephen used was working fine and gave the correct reading of the temperature.

*Does Stephen really know that the patient's temperature is 98.6 degrees, or only believe it?*

Dale-Chall Readability score: 7.0

Presented by: Ziółkowski (2015)

Truetemp Case:

One day Charles is suddenly knocked out by a falling rock, and his brain becomes re-wired so that he is always absolutely right whenever he estimates the temperature where he is. Charles is completely unaware that his brain has been altered in this way. A few weeks later, this brain re-wiring leads him to believe that it is 71 degrees in his room. Apart from his estimation, he has no other reasons to think that it is 71 degrees. In fact, it is at that time 71 degrees in his room.

*Does Charles really know that it is 71 degrees in the room, or does he only believe it?*

Dale-Chall Readability score: 6.8

Presented by: Weinberg, Nichols, & Stich (2001); Seyedsayamdost (2015); Cullen (2013)

Note: As with many of these cases, the specific wording tends to vary between studies, but this is the general form.

Truetemp Case , Community-wide Version:

The Faluki are a large but tight knit community living on a remote island. One day, a radioactive meteor strikes the island and has one significant effect on the Faluki—it changes the chemical make-up of their brains so that they are always absolutely right whenever they estimate the temperature. The Faluki are completely unaware that their brains have been altered in this way. Kal is a member of the Faluki community. A few weeks after the meteor strike, while Kal is walking along the beach, the changes in his brain lead him to believe that it is 71 degrees where he is. Apart from his estimation, he has no other reasons to think that it is 71 degrees. In fact, it is at that time exactly 71 degrees where Kal is.

*Does Kal really know that it is 71 degrees, or does he only believe it?*

Dale-Chall Readability score: 7.2

Presented by: Weinberg, Nichols, & Stich (2001)

Truetemp Case, Elders Version:

One day John is suddenly knocked out by a team of well-meaning scientists sent by the elders of his community, and his brain is re-wired so that he is always absolutely right whenever he estimates the temperature where he is. John is completely unaware that his brain has been altered in this way. A few weeks later, this brain re-wiring leads him to believe that it is 71 degrees in his room. Apart from his estimation, he has no other reasons to think that it is 71 degrees. In fact, it is at that time 71 degrees in his room.

*Does John really know that it was 71 degrees in the room, or does he only believe it?*

Dale-Chall Readability score: 6.5

Presented by: Weinberg, Nichols, & Stich (2001)

### Vegan Cookie Case:

Sharon and Mark were newlyweds who had just moved in together for the first time. They were in the process of making their new house feel like home, and adjusting to household life together. Some things were taking some getting used to, like the fact that Sharon was vegan, but Mark was not. Really it wasn't so bad, but sometimes it made grocery shopping a bit difficult.

One afternoon, Sharon was at home and wanted to make herself a snack. She really had a craving for some chocolate chip cookies. She took a cookie out of the jar, put it on a plate, and poured herself a big glass of almond milk. She was really looking forward to this cookie. However, just then she began to worry that Mark might have purchased the wrong cookies when he went shopping earlier that day. He might have purchased cookies made with eggs—or worse, lard—that would break her vegan diet.

Sharon and Mark kept things like cookies in some nice glass jars they had gotten as a wedding present, so the cookies weren't in a labeled package. She didn't want to call Mark and ask him because she didn't want to seem like a nag. So, she set about trying to determine whether the cookie was vegan. She looked in the trash and right on top she found some vegan cookie packaging that had been thrown out. Sharon looked closely at the cookie on her plate. It did look just like the photo on the packaging.

In truth, Mark did sometimes make mistakes while shopping, but this time he hadn't. The cookie was 100% vegan. However, the packaging she found was actually some old packaging from her last shopping trip. Mark had dug it up from the bottom of the trash earlier that week when he wanted to check the packaging to remember which brand she liked.

All in all, it seemed to Sharon that she had good evidence that the cookie was vegan. That was a relief, since her craving for chocolate chip cookies was



intense, and all this investigating had only made her want one more. Although she was worried earlier, now she thought the cookie was vegan. She took a bite, and then another, and another. When she was done she washed it all down with the almond milk. The cookie had hit the spot just perfectly.

*Did Sharon really know the cookie was vegan, or only believe it?*

Dale-Chall Readability score: 6.4

Presented by: Powell, Horne, Pinillos, & Holyoak (2015)

#### Watch Burglar Case:

Peter is in his locked apartment reading, and is about to have a shower. He puts his book down on the coffee table, and takes off his black plastic watch and leaves it on the coffee table. Then he goes into the bathroom. As Peter's shower begins, a burglar silently breaks into the apartment. The burglar takes Peter's black plastic watch, replaces it with an identical black plastic watch, and then leaves. Peter is still in the shower, and did not hear anything.

*Does Peter really know there is a watch on the table, or only believe it?*

Dale-Chall Readability score: 5.8

Presented by: Adleberg, Thompson, & Nahmias (2015); Starmans & Friedman (2012)

#### Yogurt Case:

Julie buys a container of yogurt at the local deli. Although, Julie is not aware of it, there is no yogurt in the container—a mixup at the factory caused the container to be filled with sour cream instead. Julie comes home, puts it her fridge, and then goes into her bedroom. Julie's neighbor Sam has been spying on her. While she is in her bedroom, he picks the lock to her apartment, and enters. He takes the

yogurt container from the fridge, and replaces it with a sealed container of yogurt from his own fridge. Then he goes back into his own apartment with Julie's yogurt container. Julie has only been in the bedroom for a few minutes, and did not hear anything.

*Does Julie really know there is yogurt in her fridge, or only believe it?*

Dale-Chall Readability score: 8.0

Presented by: Starmans & Friedman (2012)

#### Zebra Case Version 1:

Pat is at the zoo with his son, and when they come to the zebra cage, Pat points to the animal and says, "That's a zebra." Pat is right—it is a zebra. However, given the distance the spectators are from the cage, Pat would not be able to tell the difference between a real zebra and a mule that is cleverly disguised to look like a zebra. And if the animal had really been a cleverly disguised mule, Pat still would have thought that it was a zebra.

*Does Pat really know that the animal is a zebra, or does he only believe that it is?*

Dale-Chall Readability score: 1.5

Presented by: Weinberg, Nichols, & Stich (2001); Seyedsayamdost (2015)

#### Zebra Case Version 2:

Zach has an appointment with his lawyer in an office building in New York City. As he enters the lobby on the first floor, he sees something highly unexpected: a large animal with black and white stripes under a banner that says, "Pet a zebra for children's charity." In exchange for a \$10 donation to a local children's charity, you get to pet this illustrious animal. Zach quickly walks up the

stairs to his lawyer's office on the second floor. He greets the receptionist and says, "Guess what? There is a zebra down on the first floor of the building."

The people running the charity could not afford to rent a real zebra for the charity drive. So instead they hired an artist to paint black and white stripes on a mule. The animal Zach saw in the first-floor lobby was actually a cleverly disguised mule. It looks just like a zebra, but it isn't. It's a mule.

The office building that Zach is in is very large. One of the companies renting space in the building is in the business of importing exotic animals. It is illegal for them to keep these animals in an office building, but they do it anyway. They recently acquired a zebra and are keeping it well-hidden in a locked, sound-proof room on the first floor of the building.

*Does Zach really know there is a zebra on the first floor, or only believe it?*

Dale-Chall Readability score: 6.1

Presented by: Turri (2013)