Registered Replication Report: A Large Multilab Cross-Cultural Conceptual Replication of Turri, Buckwalter, & Blouw (2015)

Braeden Hall[1], Kathleen Schmidt[2], Jordan Wagge[3], Savannah C. Lewis[2,75], Sophia Weissgerber[4], Felix Kiunke[4], Gerit Pfuhl[5], Stefan M. Stieger[6], Ulrich S. Tran[7], Krystian Barzykowski[8], Natalia Bogatyreva[9], Marta Kowal[10], Karlijin Massar[11], Felizitas Pernerstofer[6], Piotr Sorokowski[12], Martin Voracek[7], Christopher R. Chartier[2], Mark J. Brandt[13], Jon E. Grahe[14], Asil A. Özdoğru[15], Michael R. Andreychik[16], Sau-Chin Chen[17], Thomas R. Evans[18], Caro Hautekiet[19], Hans IJzerman[20,76], Pavol Kačmár[21], Anthony J. Krafnick[22], Erica D. Musser[23], Evie Vergauwe[19], Kaitlyn M. Werner[24], Balazs Aczel[25], Patrícia Arriaga[26], Carlota Batres[27], Jennifer L. Beaudry[28,77], Florian Cova[19], Simona Ďurbisová[21], Leslie D. Cramblet Alvarez[29], Gilad Feldman[30], Hendrik Godbersen[31], Jaroslav Gottfried[32], Gerald J. Haeffel[33], Andree Hartanto[34], Chris Isloi[35], Joseph P. McFall[36], Marine Milyavskaya[37], David Moreau[38], Ester Nosáľová[21], Kostas Papaioannou[39], Susana Ruiz-Fernandez[31], Jana Schrötter[21], Daniel Storage[29], Kevin Vezirian[40], Leonhard Volz[41], Yanna J. Weisberg[42], Qinyu Xiao[30], Dana Awlia[2], Hannah W.

Branit[29], Megan R. Dunn[43], Agata Groyecka-Bernard[12], Ricky Haneda[14], Gabriela Kalistová[32], Julita Kielinska[8], Caroline Kolle[4], Paweł Lubomski[8], Alexys M. Miller[44], Martin J. Mækelæ[5], Mytro Pantazi[45], Rafael R. Ribeiro[26], Rob M. Ross[46], Agnieszka Sorokowska[12], Christopher L. Aberson[47], Xanthippi Alexi Vassiliou[41], Bradley J. Baker[48], Miklos Bognar[25], Chin Wen Cong[49], Alex F. Danvers[50], William E. Davis[51], Vilius Dranseika[8], Andrei Dumbravă[52,78], Harry Farmer[18], Andy P. Field[53], Patrick S. Forscher[54], Aurélien Graton[55], Nandor Hajdu[25], Peter A. Howlett[56], Radosław Kabut[8], Emmett M. Larsen[57], Sean T. H. Lee[58], Nicole Legate[43], Carmel A. Levitan[59], Neil Levy[9,46], Jackson G. Lu[60], Michael Misiak[9,12], Roxana E. Morariu[61], Jennifer Novak[4], Ekaterina Pronizius[7], Irina Prusova[62], Athulya S. Rathnayake[38], Marina O. Romanova[63], Jan P. Röer[64], Waldir M. Sampaio[65], Christoph Schild[66], Michael Schulte-Mecklenbeck[67], Ian D. Stephen[68], Peter Szecsi[25], Elizabeth Takacs[2], Julia N. Teeter[33], Elian H. Thiele-Evans[28], Julia Valeiro-Paterlini[2], Iris Vilares[69], Louise Villafana[4], Ke Wang[70], Raymond Wu[71], Sara Álvarez-Solas[72], Hannah Moshontz[73], & Erin M. Buchanan[74]

[1] Southern Illinois University Carbondale

[2] Ashland University

[3] Avila University

[4] University of Kassel

[5] UiT The Arctic University of Norway

[6] Karl Landsteiner University of Health Sciences

[7] University of Vienna

[8] Jagiellonian University

[9] University of Oxford

[10] University of Wrocław

[11] Maastricht University

[12] University of Wrocław

[13] Michigan State University

[14] Pacific Lutheran University

[15] Üsküdar University

[16] Fairfield University

[17] Tzu-Chi University

[18] University of Greenwich

[19] University of Geneva

[20] Université Grenoble Alpes

[21] Pavol Jozef Šafárik University in Košice

[22] Dominican University

[23] Florida International University

[24] University of Toronto

[25] Eötvös Loránd University

[26] ISCTE - University Institute of Lisbon

[27] Franklin and Marshall College

[28] Swinburne University of Technology

[29] University of Denver

[30] University of Hong Kong

[31] FOM University of Economics and Management

[32] Masaryk University

[33] University of Notre Dame

[34] Singapore Management University

[35] Unaffiliated Researcher, London, UK

[36] University of Rochester

[37] Carleton University

[38] University of Auckland

[39] Aristotle University of Thessaloniki

[40] Université Grenoble Alpes

[41] University of Amsterdam

[42] Linfield University

[43] Illinois Institute of Technology

[44] Adams State University

[45] Université Libre de Bruxelles

[46] Macquarie University

[47] Cal Poly Humboldt

[48] Temple University

[49] Tunku Abdul Rahman University of Management and Technology

[50] University of Arizona

[51] Wittenberg University

[52] George I.M. Georgescu Institute of Cardiovascular Diseases

[53] University of Sussex

[54] Busara Center for Behavioral Economics

[55] Université Savoie Mont-Blanc

[56] State University of New York at Fredonia

[57] Stony Brook University

[58] James Cook University

[59] Occidental College

[60] Massachusetts Institute of Technology

[61] Babeş-Bolyai University

[62] HSE University

[63] National Research University Higher School of Economics

[64] University of Witten/Herdecke

[65] Universidade Federal de São Carlos

[66] University of Siegen

[67] University of Bern

[68] Nottingham Trent University

[69] University of Minnesota

[70] Harvard University

[71] University of British Columbia

[72] Universidad Regional Amazónica Ikiam

[73] University of Wisconsin-Madison

[74] Harrisburg University of Science and Technology

[75] The University of Alabama System

[76] Institut Universitaire de France

[77] Flinders University

[78] Alexandru Ioan Cuza University

Abstract

According to the Justified True Belief account of knowledge (JTB), a person can only truly know something if they have a belief that is both justified and true (i.e., knowledge is justified true belief). This account was challenged by Gettier (1963), who argued that JTB does not explain knowledge attributions in certain situations, later called Gettier-type cases, wherein a protagonist is justified in believing something to be true, but their belief was only correct due to luck. Lay people may not attribute knowledge to protagonists with justified but only luckily true beliefs. While some research has found evidence for these so-called Gettier intuitions (e.g., Machery et al., 2017a), Turri et al. (2015) found no evidence that participants attributed knowledge in a counterfeit-object Gettier-type case differently than in a matched case of justified true belief. In a large-scale, cross-cultural conceptual replication of Turri and colleagues' (2015) Experiment 1 (*N* = 4,724) using a within-participants design and three vignettes across 19 geopolitical regions, we did find evidence for Gettier intuitions; participants were 1.86 times more likely to attribute knowledge to protagonists in standard cases of justified true belief than to protagonists in Gettier-type cases. These results suggest that Gettier intuitions may be detectable across different scenarios and cultural contexts. However, the size of the Gettier intuition effect did vary by vignette, and the Turri et al. (2015) vignette produced the smallest effect, which was similar in size to that observed in the original study. Differences across vignettes suggest epistemic intuitions may also depend on contextual factors unrelated to the criteria of knowledge, such as the characteristics of the protagonist being evaluated.

Keywords: Folk epistemology, Beliefs, Social cognition, Epistemic intuitions, Justified True Belief, Multilevel modeling, Multilab, Replication

Registered Replication Report: A Large Multilab Cross-Cultural Conceptual Replication of Turri, Buckwalter, & Blouw (2015)

The Justified True Belief (JTB) account of knowledge (or alternative versions of it) has been an important explanation of propositional knowledge in philosophical discourse for the past two millennia (e.g., Jacquette, 1996; Moser, 2002); however, some have challenged how widely accepted it has truly been (Dutant, 2015; Turri, 2016). The JTB analysis states that a claim, or proposition, is considered knowledge if it meets three conditions (Gettier, 1963). Specifically, a person (*S*) knows a proposition (*p*), if and only if:

(i) *S* believes that *p* is true,

(ii) *p* is in fact true, and

(iii) *S* is justified in believing *p* is true.

In other words, to know something, people not only must believe a claim that is indeed true; they also must have sufficient reason for believing the claim to be true. Specifically, to know something, a person must believe a true claim that was reasonably inferred from an observation or *entailed proposition* (i.e., a truth claim that is used to infer the truth of a subsequent claim). Thus, a lucky guess that happens to reflect the truth should not be considered knowledge. However, many philosophers have argued that people's *epistemic intuitions* (i.e., intuitions about knowledge) rely on more than just the presence of justified true belief. Accordingly, they have investigated the extent to which other factors, such as luck, may play a crucial role in lay epistemology.

Gettier (1963) challenged the sufficiency of the JTB account to explain propositional knowledge by presenting two strong counterexamples that are inconsistent with its predictions. These counterexamples, later referred to as Gettier-type cases, are situations in which a person has a belief that is both true and well-supported by evidence (i.e., meets all three conditions of JTB), yet that person is not judged as possessing knowledge. In many Gettier-type cases, protagonists reasonably infer a true belief (*p*) from an entailed proposition (*e*); however, in a lucky turn of events, the validity of using *e* to infer *p* is called into question, despite *p* still turning out to be true.

In one of his original counterexamples, Gettier (1963) describes a scenario in which two men, Smith and Jones, have applied to the same job at a company. Much to Smith's disappointment, the president of the company has told Smith that Jones will ultimately get the job (entailed proposition, *e1*). Smith then notices that Jones has ten coins in his pocket (entailed proposition, *e2*). Smith then infers from *e1* and *e2* the belief (*p*) that the man who gets the job, whom he assumes will be Jones, will have ten coins in his pocket. This belief is well-founded by evidence (i.e., he counted the coins in Jones' pocket himself) and, therefore, is justified. However, unexpectedly, Smith gets the job himself. Coincidentally, Smith discovers that he also has ten coins in his own pocket. Although the specifics of this outcome were not expected, his inferred belief (*p*) that the man who has ten coins in his pocket will get the job was still true. Smith reasonably inferred a true belief (*p*) from *e1* and *e2*, but neither *e1* nor *e2* actually produce the truth of *p*. Even though Smith's belief was both true and justified, Gettier argued that Smith does not have knowledge in this case–he just got lucky. Many similar scenarios (i.e., Gettier-type cases) have since been employed to demonstrate the insufficiency of JTB to fully explain knowledge attributions.[1]

[1] What makes a scenario a true Gettier-type case has been widely debated in the literature; however, for the purpose of this predominantly empirical article, we loosely refer to scenarios from this class of philosophical thought experiments as Gettier-type cases, which we operationalize for our research below.

Epistemic intuitions that prevent people from attributing knowledge to Gettier-type case protagonists, like Smith, have since been referred to as Gettier intuitions (DePaul & Ramsey, 1998; Machery et al., 2017b; Sosa, 2007). Past research has revealed some evidence that people have a universal tendency to demonstrate Gettier intuitions for some Gettier-type scenarios (e.g., Machery et al., 2017a, 2017b; Nagel, San Juan, et al., 2013). However, the extent to which people demonstrate Gettier intuitions may be influenced by other factors that have not been widely investigated. Turri et al. (2015) presented evidence that people demonstrate different epistemic intuitions for Gettier-type cases depending on how the entailed proposition (*e*) used to infer a justified true belief (*p*) is challenged, which they argued may explain the apparent inconsistencies in past work.

The present research aimed to 1) provide a robust test of Gettier intuitions for counterfeit-object Gettier-type cases, 2) explore explanations for why Gettier intuitions vary across different scenarios, and 3) explore possible cultural and demographic differences in Gettier intuitions. A secondary goal of this project was to allow psychology students to actively contribute to replication research; students engaged in data collection and other activities as part of dozens of student-lead teams across 19 geopolitical regions.

The Role of Luck in Epistemic Intuitions

Prior work suggests that people generally exhibit Gettier intuitions for at least some Gettier-type cases. Such findings indicate that people's conception of knowledge requires more than justification, truth, and belief (e.g., Machery et al., 2017a, 2017b; Nagel, San Juan, et al., 2013). However, past results have been mixed (e.g., Powell et al., 2015). In a study by Machery et al. (2017a), participants attributed knowledge to protagonists in cases of luckily true justified belief (i.e., Gettier-type cases) significantly less than in clear cases of true justified belief. Colaço et al. (2014) also found that participants were significantly less likely to attribute knowledge in a Gettier-type case than in a similarly matched knowledge control case (i.e., a clear case of justified true belief).

However, people do not demonstrate Gettier intuitions for some Gettier-type cases (i.e., intentionally replaced evidence cases; e.g., Powell et al., 2015). Starmans and Friedman (2012) found that participants tended to attribute knowledge in a "replacement-by-backup" Gettier-type case as readily as in a clear case of knowledge (Gettier intuition not demonstrated); yet Turri et al. (2015) found that participants were less likely to attribute knowledge in a replacement-by-backup Gettier-type case than in a clear case of knowledge (Gettier intuition demonstrated). Turri et al. also found that participants attributed knowledge in a "counterfeit-object" Gettier-type case no differently than in a clear case of knowledge (Gettier intuition not demonstrated); however, Powell et al. (2015) found that participants attributed knowledge less in a counterfeit-object Gettier-type case than in a clear case of knowledge (Gettier intuition demonstrated).

In the experiment replicated in the present research, Turri et al. (2015; Experiment 1) tested whether lay people demonstrate Gettier intuitions when a salient threat to the truth of a judgment fails. Turri et al. asked participants whether a protagonist in one of three conditions (i.e., a "Threat" Gettier condition, a "No Threat" knowledge condition, and a "No Detection" ignorance condition) knew or only believed a claim. In the experimental Gettier condition, participants read a story in which a protagonist named "Darrel" correctly identifies the species of an animal (i.e., target species), despite it being the only animal of that species among many animals of a different, almost identical species (i.e., counterfeit species). Participants in the other two conditions read the same story with slight changes: In the knowledge control version, the story never mentions the other identical species (i.e., no counterfeit), and in the ignorance control version, the protagonist incorrectly identifies the counterfeit species as the target species. Turri et al. then compared the rate of knowledge attributions between participants in the Gettier condition and participants in the two control conditions. They found no evidence of Gettier intuitions; participants in the Gettier condition attributed knowledge at rates no different from those in the knowledge control condition [$\chi^2(1, N = 98) = 2.63$, Fisher's exact $p = .164$, Cramér's $V = .164$; Gettier intuition not demonstrated]. These findings suggest that luckily true justified beliefs may be consistent with lay people's conception of knowledge under certain conditions and highlight the need for further research on epistemic intuitions in Gettier-type cases.

The average size of Gettier intuition effects, and the conditions under which they emerge, is currently unknown. According to Turri (2016), knowledge attribution rates for different Gettier-type cases vary from lower than 20% (Gettier intuition demonstrated) to higher than 80% (Gettier intuition not demonstrated); although, the sources of these estimates are unclear. Such inconsistencies in knowledge attribution rates are perhaps due to two major reasons: (1) people's epistemic intuitions, which lead them to make different judgments about various types of Gettier-type cases based on the characterization of the luckily true justified belief and (2) variation in experimental designs, including differences in matched controls and some possibly underpowered samples (see Colaço et al., 2014; Machery et al., 2017b; Nagel, Mar, et al., 2013; Nagel, San Juan, et al., 2013; Powell et al., 2015; Starmans & Friedman, 2012; Turri et al., 2015; Weinberg et al., 2001).

Although the literature on epistemic intuitions has demonstrated varying attribution rates across different types of Gettier-type cases, Powell et al. (2015) and Nagel, San Juan, et al. (2013) provide evidence for Gettier intuitions using counterfeit-object Gettier-type cases. Unlike Turri et al. (2015), Nagel, San Juan, et al. (2013) found that participants were more likely to attribute knowledge to a protagonist in a standard justified true belief condition than a protagonist in a Gettier condition. In reply, Starmans and Friedman (2013) argued that Nagel, San Juan, et al. employed a questioning method that biased participants to deny knowledge, did not properly evaluate the responses of participants who may have attributed knowledge to protagonists in Gettier-type cases, misconstrued the distinction between "apparent" and "authentic" evidence, and used scenarios that did not feature the structure that characterizes most Gettier-type cases. Starmans and Friedman concluded that Nagel, San Juan, et al.'s findings are fully compatible with the claim that lay people attribute knowledge in Gettier-type cases (Gettier intuition not demonstrated; cf. Nagel, Mar, et al., 2013).

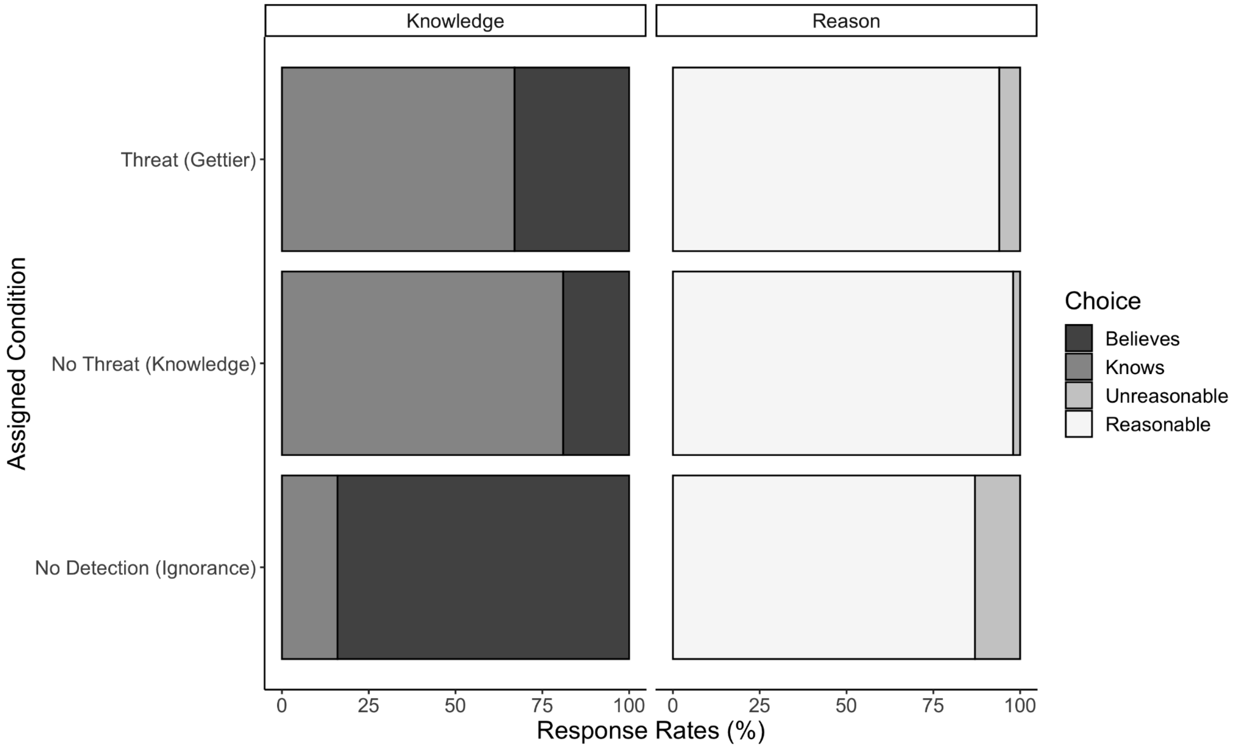
The Current Study

Some previous research suggests that lay people may be more likely to attribute knowledge to protagonists who have non-lucky justified true beliefs than to protagonists who have justified true beliefs due to luck alone, thus demonstrating Gettier intuitions (e.g., Machery et al., 2017a, 2017b; Nagel, San Juan, et al., 2013). However, other investigations have found no differences in knowledge attributions between these conditions (e.g., Starmans & Friedman, 2012; Turri et al., 2015). Because of such inconsistencies in the literature, we sought to estimate the prevalence of Gettier intuitions in a large, highly powered, and international conceptual replication of Turri et al.'s (2015) Experiment 1. In this study, we examined one subset of Gettier-type cases, counterfeit-object cases, using a variety of vignettes, carefully matched controls, and a large cross-cultural sample. Like Turri et al.'s original experiment, the current study explored the frequency of knowledge attribution in response to a protagonist making a correct inference from a false belief.

First, we tested whether participants attributed knowledge to a protagonist differently across three conditions: when their belief is justified and true (i.e., in the “No Threat” or knowledge condition), when the protagonist’s belief is justified but only true because of luck (i.e., in the “Threat” or Gettier condition), and when the protagonist’s justified belief is false (i.e., in the “No Detection” or ignorance condition). Following the results of Turri et al. (2015), we predicted that the Gettier condition would produce knowledge attributions at rates no different from the knowledge condition but more frequent than the ignorance condition. Second, we compared participant ratings of the belief’s reasonableness by condition to see if, like Turri et al., we would find no condition differences in participant perceptions of what was reasonable for the protagonist to believe. See Figure 1 for the original knowledge attribution and reasonableness results. We also attempted to replicate Turri and colleagues’ (2015) findings that participants were more likely than chance to attribute knowledge to protagonists in the No Threat (i.e., knowledge) condition ($p < .001$) and in the Threat (i.e., Gettier) condition ($p < .001$) but less likely than chance to attribute knowledge in the No Detection (i.e., ignorance) condition ($p = .021$). Finally, to increase the contribution of our replication, we tested the extent to which Turri et al.’s findings generalize across different data collections sites and vignettes.

Figure 1

Results of Turri et al. (2015), Experiment 1



Differences from Turri et al. (2015)

Past experimental philosophy research provides several methodological explanations for inconsistencies in Gettier intuition research, such as design, measurement, and culture. We modified the original Turri et al. (2015) experiment to address these concerns.

Design considerations. The consensus for explaining inconsistencies in Gettier intuition research is that the epistemological structure of Gettier-type cases varies depending on the tested vignette or case type (Turri, 2016). The two original counterexamples Gettier used in his 1963 paper each describe a protagonist who forms an initially justified but false belief from which a true claim is then inferred. Some philosophers now use the term “Gettier case” (or Gettier-type case) to refer to any instance that is intended to illustrate the non-equivalence of justified true belief and knowledge, wherein a given justified true belief is supposed to be viewed as not being consistent with knowledge (Nagel, San Juan, et al., 2013). Alternatively, others have used the term more specifically to denote cases of the particular inference-from-false-belief type structure featured in Gettier's original article, regardless of whether the case itself is viewed as consistent with knowledge (e.g., Weatherson, 2013). We do not define Gettier-type cases as instances that are intended to show a disparity between justified true belief and knowledge, as Nagel, San Juan, et al. (2013) suggested. Instead, we adopted the latter interpretation by defining Gettier-type cases as scenarios with the structure featured in Gettier's original article, which we used to guide our selection of additional related Gettier-type cases to test.

Ignoring the stimulus variation present in the experimental philosophy literature would limit the generalizability of our results (Nagel, San Juan, et al., 2013; Starmans & Friedman, 2012; see also Judd et al., 2012; Yarkoni, 2022). Thus, we attempted to conceptually replicate the original Turri et al. (2015) experiment using additional counterfeit-object Gettier vignettes from the literature (i.e., "Fake Barn" vignette from Colaço et al., 2014; "Diamond" vignette from Nagel, San Juan, et al., 2013). In these vignettes, a protagonist makes a true inference from a false belief by unknowingly and luckily choosing a true, genuine object among many convincing counterfeits. Doing so allowed us to test the generalizability of Turri and colleagues' (2015) Experiment 1 "Darrel" manipulation to other similar counterfeit-object cases while reducing stimulus sampling error. We decided to test these different vignettes using a mixed design rather than a between-participants design. Participants were randomly assigned without replacement to each condition and each vignette, resulting in each participant being presented with three vignette/condition combinations. This approach allowed us to parse out the within-participants variation, thereby increasing the statistical power of our analyses to detect and estimate the Gettier intuition effect.

Measurement considerations. Turri et al. (2015) used a binary measure to assess knowledge attribution. However, in personal correspondence (J. Turri, personal communication, March 10, 2018), Turri stated that participants in knowledge control and the Gettier condition may not have differed in their knowledge attributions in the to be replicated study due to the study's underpowered sample size and the binary format of the knowledge probe. If lay people evaluate the knowledge of others along a spectrum, then employing a more scaled measure may reveal differences that could be missed by a dichotomous measure. Subsequent research by one of the original authors measured knowledge with a 7-point Likert-type scale on which participants rated their agreement with a statement claiming a protagonist knew a given proposition (Turri, 2016; Study 2). Although this study used a slightly different vignette than Turri et al.'s Experiment 1, Turri (2016) found a sizable difference (*d* = 0.73) in participant knowledge attributions between a "Threat" (i.e., Gettier) condition and an appropriately matched knowledge control condition.[2] Potentially, the use of a scaled measure allowed for the detection of the Gettier intuition effect. In the present research, we employed a visual analogue scale (VAS) ranging from 0 to 100 in lieu of the original binary (i.e., knows/only believes) response variable. The VAS may be as efficacious as a Likert-type response scale and provides more fine-grained data for analysis via parametric statistics than alternatives by allowing for more variability in responding (Bishop & Herron, 2015). Although using a VAS departs from the

[2] The "No Threat" (i.e., knowledge control) and "Threat" (i.e., Gettier) conditions were structurally similar to the conditions used in the replicated study. Both studies featured protagonists in the woods trying to identify an animal. In the "Threat" conditions, the protagonist identifies the animal correctly but only because of some kind of situational luck.

original study, and from how these kinds of judgments are typically made in everyday life, our pretest using a VAS found that participants responded to the control conditions in the expected way with this measure (i.e., knowledge controls and ignorance controls demonstrated paradigmatic rates; see https://osf.io/3ygsk/).

Another addition to our replication was the inclusion of an exploratory knowledge probe. Differences in knowledge attribution may depend on how participants are asked whether a target has knowledge (e.g., Nagel, San Juan, et al., 2013). To check for these differences in knowledge attribution based on the form of the knowledge question, we asked an exploratory binary knowledge attribution question after the primary knowledge attribution question. We also added an exploratory item to assess perceptions of luck and ability that may moderate knowledge attributions in response to Gettier-type cases (e.g., Turri, 2016). See the Materials and Measures section below for details.

Cultural considerations. Researchers have examined potential cultural sources of variation in knowledge attribution (e.g., Buckwalter & Stich, 2010; Kim & Yuan, 2015; Machery et al., 2017a, 2017b; Nagel, San Juan, et al., 2013; Nichols et al., 2003; Seyedsayamdost, 2015; Turri, 2013; Turri et al., 2015; Weinberg et al., 2001). For example, Weinberg et al. (2001) reported evidence that participants with Western cultural backgrounds demonstrate Gettier intuitions more often than participants with Eastern cultural backgrounds. However, this preliminary study was underpowered and lacked control conditions; subsequent cross-cultural studies (that also lacked matched controls) found no such cultural differences (e.g., Machery et al., 2017a, 2017b; Seyedsayamdost, 2015). In one of the largest of these cross-cultural studies, Machery and colleagues (2017a) provided evidence that people exhibit Gettier intuitions across quite different cultures and languages (i.e., USA, Brazil, India, and Japan); they argued that humans have a "species-typical core folk epistemology" wherein justification, truth, and belief are insufficient for knowledge attribution (p. 12).

Comparisons among these past findings are difficult due to the use of different control conditions that varied in how closely matched they were to the experimental Gettier condition. While more recent studies have used both knowledge and ignorance control conditions in which participants are exposed to paradigmatic cases of knowledge and ignorance, respectively, most cross-cultural studies have not used closely matched control stimuli (e.g., Kim & Yuan, 2015; Machery et al., 2017a, 2017b; Seyedsayamdost, 2015). For example, Machery and colleagues (2017a) used a between-participants design with entirely different vignettes and different protagonists for each condition. By contrast, Turri et al. (2015) used slight variations of the same vignette for each condition. Because the versions of the Darrel vignette used in Turri et al. differ only in the words necessary to alter the condition of the protagonist's belief, we also ensured that the two added vignettes (i.e., the Fake Barn Gerald vignette and the Diamond Emma vignette) were implemented with closely matched control conditions. See Appendix B for full details.

Pedagogical Goals

A second aim of this project was to provide psychology students across the globe with the opportunity to contribute to a rigorous large-scale research study. We implemented the model of the Collaborative Replications and Education Project (CREP; Grahe et al., 2014; Wagge et al., 2019) and initiated a collaboration between the CREP and the Psychological Science Accelerator (PSA; Moshontz et al., 2018). The purpose of the CREP is to provide experiential learning opportunities for psychology students while addressing the need for direct replication work in the field of psychology by using the collective power of student research projects. The PSA is an international network of collaborators with a mission to expedite the accumulation of reliable and generalizable evidence in psychological science (Moshontz et al., 2018). The CREP and PSA partnership involved the CREP selecting a study, developing materials, and overseeing the quality of the replications using standard CREP procedures, while using the existing PSA network to increase participation among labs. Additionally, the PSA's extensive network of experts has supported lab recruitment, translations, data management, and navigating international collaborative research.

While both the CREP and the PSA have been successful models of multisite collaboration, this project was neither solely a CREP study nor solely a PSA study. The study differed from the typical CREP project in the following ways: (1) it was not a direct replication; (2) it involved a Registered Report; (3) almost all of the data collection was centralized; and (4) students were encouraged but not required to conduct site-level data analysis in order to earn a CREP completion certificate. The study also differed from the typical PSA project in the following ways: (1) it had significant pedagogical goals; (2) some data was collected independently by labs rather than with a centralized survey; and (3) teams were more autonomous in how they implemented the project. At times, methodological decisions pitted scientific priorities against pedagogical priorities, and pedagogy was prioritized. For example, we allowed students to collect data via Qualtrics surveys that they had created themselves, which allowed for more autonomy and opportunities for students to develop skills but also resulted in some data loss and processing difficulties (see Methods and Appendix A).

Summary

Previous research has produced mixed evidence regarding the presence and size of Gettier intuition effects. Some of this variation may be explained by differences in the design, measurement, and cultural contexts found across previous investigations. Using counterfeit-object Gettier-type cases, we sought to estimate the effect size of Gettier intuitions across a variety of geo-political contexts while attempting to address methodological concerns (i.e., measurement sensitivity, lack of matched controls, and stimulus variation). Our results provided evidence regarding the prevalence of Gettier intuitions among lay participants, the extent to which Gettier intuitions are shared across cultures, and the stability of Gettier intuitions across similar scenarios with different protagonists in different contexts.

Disclosures

Preregistration

This study was provisionally accepted as a Registered Replication Report and subsequently preregistered on the Open Science Framework (OSF; see https://osf.io/4bfs7).

Data, Materials, and Online Resources

Study materials, de-identified raw data, de-identified data with exclusions, and analysis code and output are available on our master OSF page (https://osf.io/n5b3w/). Many project teams also posted data on their team's OSF page linked to our master page.

Reporting

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (see Simmons et al., 2011).

Ethical Approval

All contributing project teams were required to submit their local institutional ethics approval (if applicable) prior to data collection as part of their pre-registration and CREP review process.

Methods

Deviations from Provisionally Accepted Protocol

The protocol for this study was accepted as a Stage 1 Registered Replication Report (https://osf.io/37p8t/; see also Appendix A). In this section, we describe the method as implemented and deviations from the protocol, including minor adjustments to language, corrections of factual inaccuracies, and methodological alterations. The primary deviations from the approved protocol, albeit minor, consisted of changes to study procedure and the analysis plan due to error and adaptations required for valid statistical inference. As detailed below, we changed the methodology according to how surveys were programmed and implemented, how we measured luck attribution, how we measured race/ethnicity, and how we determined the inclusion of data from the student-led teams. We additionally chose to drop two of the planned covariates, whether the study was conducted individually versus in a group setting and in-person versus online, because they were unusable.[3] A number of aspects were not sufficiently described in the original protocol; we therefore clarified the analysis plan in terms of exclusion criteria and data assumption checking procedures.

Project Teams

Each student-led project team prepared a study protocol for approval by a CREP reviewer to ensure quality control. Teams could not contribute to data collection until their protocol was approved. For more information about this process and detailed descriptions of logistical

[3] The COVID-19 pandemic changed, and significantly limited, how students could carry out their replication studies. After it began, our data collection was shifted to almost entirely online (and individual) participation. As can be seen in Table 1, most sites had online and individual sessions, some of the sites had both session types for one or both of the two variables, and some sites were missing documentation. Thus, using the covariates as intended would have been impossible.

considerations, see Appendix A.[4] In total, 65 student-led teams (i.e., unique teams with OSF pages) signed up to collect data for this project, and 51 student-led project teams were approved to begin data collection using CREP procedure guidelines. Only 47 of these teams contributed to the full dataset, which represented 38 data collection sites. See Table 1 for a summary of the sites and their data collection features. Teams were not included in the full dataset either because they did not collect any data (e.g., due to campus closures during the COVID-19 pandemic) or because the data they collected were unusable for analyses (e.g., vignettes were not properly randomized). After applying the participant level exclusions described below, the final dataset included 45 student-led project teams across 37 data collection sites. Of those 45 teams, 22 received CREP completion certificates for completing all pedagogical tasks (e.g., site level analyses). While we initially planned to include only the data from teams that received completion certificates, we decided to include all usable data from teams that were approved to start data collection (see Analytic Approach).

[4] The Stage 1 registered report manuscript included sections that described the recruitment and approval of collaborators who would collect data. We have restructured the Method section to more closely resemble that of a typical empirical article. The original text, updated to reflect the study's completion, can be found in Appendix A.

Table 1

Characteristics of Data Collection Sites

Language	Geopolitical Region	# Teams	Full *N*	Final *N*	Sample	In person?	In a group?	Compensation	Site-level Analysis
Chinese (Traditional)	Taiwan	1	452	89	Undergraduates	No	No	Money	ANOVA (VAS) + Chi-square (binary)
English	Australia	1	215	122	Undergraduates	No	No	Credit	None
		1*	165	119	Undergraduates	No	No	Credit	None
	Canada	1	551	258	Undergraduates	No	No	Credit	None
	United Kingdom	2	340	132	Undergraduates	No	No	Credit	Friedman
	Greece	1	98	52	Both	No	No	None	None
	New Zealand	1	58	42	Undergraduates	Yes	No	Lottery	ANOVA
	Singapore	1	78	52	Undergraduates	Yes	No	Credit	None
	United States	1	124	57	Undergraduates	No	No	Credit	ANOVA
		1	387	221	Undergraduates	Yes	Unclear	Credit	Linear Mixed Model
		2	402	201	Undergraduates	No	No	Credit	ANOVA
		1	91	48	Undergraduates	Both	No	Credit	None
		1	164	93	Undergraduates	No	No	Credit	ANOVA
		1	64	43	Undergraduates	Yes	Unclear	Credit	ANOVA
		1	187	78	Undergraduates	No	No	Credit	Descriptives only
		1	129	56	Undergraduates	No	No	Credit	ANOVA
		1*	510	356	Community	No	No	Money	Linear Mixed Model
		1	213	103	Undergraduates	No	No	Credit	ANOVA

Language	Geopolitical Region	# Teams	Full *N*	Final *N*	Sample	In person?	In a group?	Compensation	Site-level Analysis
		1	135	63	Both	No	No	Students: Credit; Community: None	None
		1	162	93	Undergraduates	No	No	Credit	None
		1	103	36	Undergraduates	No	No	Credit	None
		1	112	56	Undergraduates	No	No	Credit	Descriptives only
		1	95	54	Undergraduates	No	No	Credit	None
		1	8	NA	Undergraduates	Unclear	Unclear	Unclear	None
French	Switzerland	1	58	28	Undergraduates & Community	Both	Both	Undergraduates: Credit Community: Money	Bayesian ANOVA
German	Austria	1	159	76	Both	No	No	None	None
		1	143	75	Both	No	No	None	None
	Germany	7	1102	588	Undergraduates	Both	Both	Varied across teams (sweets, money, credit)	One team used McNemar for binary and Quade test for VAS; otherwise none
		1	332	184	Undergraduates	No	No	Credit	None
Hungarian	Hungary	1	658	449	Undergraduates	Both	Both	Credit	None
Norwegian	Norway	1	147	76	Both	No	No	Students: Sweets; Community: None	None
Polish	Poland	1	179	72	Both	Both	Both	Students: Credit Community: None	MANOVA
		1	243	121	Both	No	No	None	None
Portuguese	Portugal	1*	149	81	Both	No	No	Students: Credit; Community: None	Chi square & binomial tests

Language	Geopolitical Region	# Teams	Full *N*	Final *N*	Sample	In person?	In a group?	Compensation	Site-level Analysis
Romanian	Romania	1	661	371	Undergraduates	No	No	Credit	None
Russian	Russia	1	233	99	Both	No	No	None	None
Slovak	Slovakia	1	229	105	Both	No	No	None	ANOVA
Turkish	Turkey	1	304	77	Undergraduates	No	No	Credit	None
	Total	46	9440	4826					

Note. Full and Final *N* indicate sample size before and after exclusions. Data collection context variables were gleaned from OSF page documentation and confirmed by the team when possible. “Unclear” indicates lack of documentation. Site-level analyses were conducted independently and not included in the analyses presented here. See the team OSF pages for details and results.

*Team collected data using Qualtrics instead of SoSciSurvey

Participants

In the analysis sample (i.e., after the exclusions described below), participants were 4,826 adults recruited to participate by student researchers at 37 data collection sites in various geopolitical contexts across geographical regions (i.e., Northern America, Eastern Europe, Western Europe, Northern Europe, Southern Europe, Australia and New Zealand, Western Asia, Southeastern Asia, Eastern Asia). See Table 2 for sample sizes by geopolitical region. Data collection took place between January 1, 2019, and June 1, 2021.[5] Data collection sites contributed a median of 81 participants to the analysis sample (*min* = 28, *max* = 588); 6 sites collected fewer participants than the target of 50. On average, participants were young (M_{age} = 24.84, *SD* = 9.91, *n* = 4,826) and had completed some college as measured by years of education ($M_{education}$ = 13.84, *SD* = 2.59, *n* = 4,771).[6] Most participants (70.37%; *n* = 3,396) identified as White.[7] Over half of participants identified as female (70.56%; *n* = 3,405) and most other participants identified as male (29.01%; n_{male} = 1400; $n_{neither}$ = 21). The plurality of participants completed the survey in English (47.53%; *n* = 2294). Participation details, like compensation and the sampled population, varied by data collection site. See Table 3 for summary.

[5] In the approved protocol, we described a plan for data collection whereby each lab preregistered a target sample size of 50-100 and stopped collecting data on April 1, 2020, or once all contributors reached their preregistered target sample size. Due to the COVID-19 pandemic, this plan was not followed. The deadline for data collection was extended to June 1, 2021. Many data collection sites stopped collecting data earlier.

[6] There may be measurement error in participants' reported years of education. Less than the equivalent of a high school diploma was reported by 620 participants, 52 of whom reported one year of education.

[7] While we planned to measure participants' racial and ethnic identities using an open-ended response, racial and ethnic identity was measured using non-exclusive categories with an open-ended fill-in option for reasons that were not documented. Student research teams designed different response options tailored to their geographic region (see all variations in Appendix C). All data collection sites allowed people to select multiple racial and ethnic identities, and all asked whether participants identified as White (either "White/European", "White/European descent", or "European descent").

Table 2

Number and Percentage of Participants in the Analysis Dataset (after Exclusions) by Geopolitical Region

Geopolitical Region	*n*	% of total
United States	1558	32.28
Germany	772	16.00
Hungary	449	9.30
Romania	371	7.69
Canada	258	5.35
Australia	241	4.99
Poland	193	4.00
Austria	151	3.13
United Kingdom	132	2.74
Slovakia	105	2.18
Russia	99	2.05
Taiwan	89	1.84
Portugal	81	1.68
Turkey	77	1.60
Norway	76	1.57
Greece	52	1.08
Singapore	52	1.08
New Zealand	42	0.87
Switzerland	28	0.58

Note. Geopolitical region refers to the location of the data collection site except for one team that collected data through Amazon Mechanical Turk (MTurk) in another geopolitical region (i.e., the United States). For all other data collection sites, participants were recruited from the geopolitical region of the site.

Table 3

Number and Percentage of Participants by Data Collection Context Variables

Variable	*n*	% of total
Compensated for participation	3533	73.21
Recruited through mTurk	356	7.38
Completed the centralized survey	4270	88.48

Note. Variables are not exclusive. Information about the compensation method was obtained by examining each student-led team's IRB approval, confirming with the students or PIs at each site, and making inferences based on the data collection site specific surveys when neither source was available. Three data collection teams included in analyses used Qualtrics to distribute their surveys instead of the centralized survey programmed in SoSciSurvey.

Exclusions

Of the 9,440 participants who completed the survey, data from 48.88% (*n* = 4,614) were excluded from the analytic sample. Of this total, 2,187 participants (23.17%) were flagged for exclusion based on multiple criteria. All listed exclusions were preregistered with one exception (i.e., maximum age).[8] Participants were excluded for the following reasons:

- Age: the participant did not provide an age, listed an age greater than or equal to 100, or was not the age of majority of their geopolitical region, operationalized as at least 18 in all regions except Taiwan, where the age of majority is 20 (total excluded: *n* = 2,118; missing: *n* = 2,040; 22.44% of participants met this exclusion criterion).
- Prior participation: the participant had taken part in a previous version of this study or in another contributors' replication of the same study (*n* = 238; 2.52% of participants met this exclusion criterion).

[8] We did not preregister the exclusion of people who reported their age as over 100; only 7 people were excluded on the basis of this criteria alone (i.e., they did not meet any other exclusion criteria). These responses may have been errors in data entry or unlabeled test responses.

- Comprehension: the participant failed to answer all three of the vignette comprehension questions correctly (e.g., did not correctly identify whether Darrel was looking at a squirrel or a prairie dog; total excluded: *n* = 4,376; missing: *n* = 1,490; 46.36% of participants met this exclusion criterion).[9] See Table 4 for rates of correct responses by vignette and condition combination.
- Knowledge of hypothesis: the participant correctly and explicitly articulated knowledge of the specific hypotheses or specific conditions of this study when asked what they thought the study hypothesis was (*n* = 203; 2.15% of participants met this exclusion criterion).
- Language proficiency: the participant reported their understanding of the language the survey was presented in as "not well" or "not well at all" (total excluded: *n* = 2,093; missing: *n* = 2,003; 22.17% of participants met this exclusion criterion; see Vickstrom et al., 2015 for criteria).

See Materials and Measures section below for item details. The rate at which participants were excluded due to failed comprehension in the present study (46%) was consistent with prior cross-cultural Gettier intuition research (e.g., rates between 21% [Machery et al., 2017b] and 47% [Machery et al., 2017a]). Across Gettier intuition studies more broadly, such exclusions have rarely had an impact on results (for review, see Popiel, 2016).

[9] Turri et al. (2015) used the same type of question for the same purpose and excluded 15 of 135 participants on this basis.

Table 4

Comprehension Question Correct Answer Rates by Condition and Vignette Combination

	Gettier		Ignorance		Knowledge	
	Total	Correct	Total	Correct	Total	Correct
Darrel	2821	1986 (70.40%)	3153	2119 (67.20%)	2972	2174 (73.15%)
Emma	2982	2009 (67.37%)	3034	2104 (69.35%)	2930	2085 (71.16%)
Gerald	3143	1942 (61.79%)	2759	2001 (72.53%)	3044	2035 (66.85%)
Missing across vignettes and condition	494		494		494	

Note. Participants were excluded from analyses if they incorrectly answered any of the comprehension questions.

Power Analysis

We conducted an *a priori* power analysis, using the *powerCurve* function in the *simr* package (Green & MacLeod, 2016) in *R* to estimate the sample size required to detect an effect of knowledge condition on participants' knowledge attributions with 90% power at $\alpha = .05$.[10] To estimate the effect size, we considered (1) the effects observed in our pilot-test data (difference between Gettier and knowledge, $\beta = 0.32$; difference between Gettier and ignorance, $\beta = -0.44$), (2) both the difference between the Gettier condition and knowledge condition (Cramér's $V = .509$) and the small non-significant difference between the Gettier condition and ignorance condition (Cramér's $V = .16$) from Experiment 1 of Turri et al. (2015), and (3) the small effects sometimes found in the literature (e.g., Machery et al., 2017a). To be conservative, we selected a

[10] The approved protocol described a power analysis conducted prior to data collection. The text from the original protocol is reproduced in full in Appendix A and is summarized here.

standardized fixed effect within the multilevel model analysis described below of .1 for our power analyses.

The model tested included random intercepts for data collection site, vignette, and participants, such that vignettes were nested within participants who were nested within sites. We simulated data using a standardized fixed effect regression parameter of .1. In these simulations, the number of participants per site was allowed to vary, but the number of vignettes (3) and the number of collection sites (9) were held constant. Results suggested that at least 32 participants per data collection site (i.e., 288 total participants; 864 total observations) would be necessary to detect the identified fixed effect regression parameter (.1) 90% of the time with an alpha of .05. Considering the potential for attrition (e.g., due to lack of comprehension) and effect size heterogeneity between data collection sites (Kenny & Judd, 2019), we set a target sample size of 50 participants per data collection site. Of the 46 data collection sites included in analyses, 45 met this target prior to exclusions, and 40 met the target after exclusions.

Materials and Measures

As described in the approved protocol, we planned to collect all data using a single SocSciSurvey survey programmed to accommodate lab-specific variations. However, eight student-lead teams used Qualtrics surveys programmed by student researchers; some Qualtrics teams used versions created by other Qualtrics teams. The majority of the data collected via Qualtrics was not included in the full data set due to logistical challenges (e.g., no access to raw survey data); only three of the teams included in the analysis dataset used Qualtrics surveys (*n* =

556 after exclusions).[11] All materials used in this replication are available in Appendix B and at https://osf.io/n5b3w.

Vignettes

In addition to the "Squirrel/Darrel" vignette from Turri et al. (2015), two vignettes were selected on the basis of their similarity to the original vignette, their quality, and their prevalence in the literature: the "Fake Barn/Gerald" vignette (Colaço et al., 2014; altered to more closely match the "Squirrel/Darrel" vignette), and the "Diamond/Emma" vignette (Nagel, San Juan, et al., 2013). The vignettes as administered in this study are reported in full in Appendix B. The vignettes were pretested to ensure they effectively manipulated the target construct and produced sufficient participant comprehension (see https://osf.io/3ygsk/). Four student-lead teams participated in an optional extension that included a fourth vignette after the main study protocol to test the effects of perceived expertise on Gettier intuitions (see Larkin & Andreychik, 2019). However, we did not use the data from this extension in any of the analyses reported in this paper.

For each vignette, participants were randomly assigned without replacement to one of three conditions: a Gettier-type condition in which the vignette subject correctly identified the target but not due to the reason they thought it to be true (i.e., the "Threat" condition in Turri et al., 2015), a knowledge control condition in which the subject correctly identified the target due to their knowledge (i.e., the "No Threat" condition in Turri et al.), and an ignorance control

[11] A set of multilevel models examined if the data source (Qualtrics versus SoSociSurvey) interacted with experimental condition in predicting knowledge, reasonableness, and luck judgments. No interaction was found in these analyses, which can be viewed at https://osf.io/nvfbm. Therefore, all data were combined into one large dataset after matching variables.

condition in which the protagonist incorrectly identified the target (i.e., the "No Detection" condition in Turri et al.).

Dependent Measures

After each vignette, two primary and two exploratory dependent variables were measured. In line with the approved protocol, all student-led teams included the default visual analog scale ranging from 0 to 100 for three of these variables (i.e., knowledge attributions, reasonableness judgments, and attributions to luck vs. ability). However, six teams also participated in an optional extension that randomly assigned participants to take the study with either entirely continuous scale measures or entirely binary choice measures for these variables.[12] Overall, for each of the three measures, 86.52% of responses used in analyses were originally measured on the continuous scale. Exact question text is reported in Appendix B.

Knowledge Attributions. Participants were asked whether the protagonist believes or knows the stated proposition.

Reasonableness Judgments. Participants were asked to rate the extent to which the protagonist's belief was unreasonable or reasonable.

Luck/Ability Attributions. For this exploratory measure, participants were asked two questions relevant for evaluating their attributions of outcomes to luck or ability. First, participants were asked whether or not the protagonist got the "right" or "wrong" answer. Then, participants were asked whether the protagonist's "right" or "wrong" answer was due to their

[12] Teams that participated in this extension were required to collect twice as many participants (n > 100; half in the continuous condition and half in the binary condition) so that they could meet the sample size requirement (n = 50) for participants using only the pre-approved continuous measure. However, because we converted all continuous responses to binary responses (see Analytic Approach section below for more details), the binary responses collected using this extension were also included with the converted binary responses in analyses.

ability/inability or their good luck/bad luck on one of the two scales.[13] If participants selected the incorrect answer to the first part of the question, they were subsequently excluded from the luck attribution analyses because their response indicated that they did not comprehend whether or not the protagonist held the given true belief.

Alternative Knowledge Attribution. In addition, participants were asked a binary alternative knowledge probe in which participants chose whether the protagonist either knew what the target of identification was or felt like they knew what the target was but did not actually know. For example, participants were asked, “In your view, which of the following sentences better describes Darrel’s situation?” after the Darrel vignette. Participants could then select one of two response options: “Darrel knows that the animal he saw is a red speckled ground squirrel.” or “Darrel feels like he knows that the animal he saw is a red speckled ground squirrel, but he doesn’t actually know that it is.”

Demographics and Participation Characteristics

Participants were asked to report their age, gender, geopolitical region (i.e., “What country do you currently live in?” and “What is your country of birth?”), the number of years they had attended school, and their race or ethnicity. Because of differences in how student-led teams measured these items, we matched item answers across different implementations of the survey. Participants also completed a 12-question study experience questionnaire that was not used in analyses (see Appendix C).

Education Level. All participants were asked a question about their education. Participants who completed the study in SocSciSurvey were asked about the number of years

[13] The two-part luck/ability attribution was planned as a single item with two responses presented on a single screen. The presentation of the measure was modified to reduce participant confusion by splitting the two parts across two items on separate screens.

they had been in school (truncated at 18). Participants who completed the survey in Qualtrics were asked about their educational attainment. Education (in years) was imputed for participants who reported their educational attainment from these three sites (*n* = 553).[14] The years of education for these sites was also truncated to match how this item was measured in SocSciSurvey, such that any value above 17 was recoded as 18.

Compensation. Participants were asked whether or not they were compensated for their participation (i.e., “Will you receive any kind of compensation or reward for taking part in this study?”), and indicated the type of compensation (e.g., the number of course credits, the amount of money). Some student-led teams opted not to include this question in their survey because all participants were compensated the same way. The method of compensation described in the site’s approved IRB protocol was imputed for those missing responses. Among participants who were asked about their compensation, responses were sometimes missing or discrepant with the documented method of compensation. For student-led teams where fewer than 50% of participants in the final dataset agreed on a method of compensation, the method of compensation described in the data collection site’s approved IRB protocol was imputed for all participants if a single method of compensation was described.

Comprehension and Language Proficiency. Participants were asked to indicate the true correct answer for each vignette as a comprehension check that was used for listwise exclusions. Participants were also asked to rate their proficiency for the survey language. The original paper asked participants whether they were native English speakers but did not seem to exclude

[14] For participants from the United States, less than a high school education was coded as 10 years, a high school diploma was coded as 12 years, some college or a 2-year college degree was coded as 14 years, a 4-year college degree was coded as 16 years, a master’s was coded as 18 years, and a doctorate or professional degree was coded as 20 years. For participants from Portugal, the labels and coding were the same except that a 3-year college degree was coded as 15 years and a doctorate degree was coded to 21 years.

participants on this basis. Given that the tasks in the present study were highly dependent on language comprehension and proficiency, and that participants had a 12.5% chance (i.e., 1 in 8) of passing all three comprehension questions based on guesses, we decided an additional check of self-reported language proficiency would be helpful in excluding participants who did not or may not have understood the task completely.

Prior Participation and Knowledge of Study. We also asked participants to describe what they thought the hypothesis of the study was (used for exclusions), to provide their impression of study materials (not used in any analyses), and whether they had participated in a similar study (used for exclusions). The original study did not contain these three questions, but the researchers excluded Amazon Mechanical Turk (MTurk) workers if they had already participated. Evaluating the hypothesis and prior participation exclusion criterion required subjective judgments about open-ended responses. Each non-missing observation was evaluated by three raters who spoke the language of the provided response. These three raters did not translate responses, but instead directly evaluated responses with respect to the exclusion criteria. Responses marked "yes" (i.e., meets criteria) were assigned 2 points, responses marked "maybe" (i.e., may meet criteria) were assigned 1 point, and responses marked as "no" (i.e., does not meet criteria) were assigned 0 points. After summing points for each response across the three raters, we excluded cases with 4 or more points on either response. See Appendix D for the instructions given to raters and http://osf.io/gs29c for the ratings data. Responses identified by raters as test cases (e.g., "TEST") were excluded (study purpose: *n* = 222; previously participated: *n* = 170).[15]

[15] Data collection sites were not given instructions about avoiding or clearly identifying test responses. At many data collection sites, the students and other researchers executing the study tested their survey link multiple times (e.g., as inferred by responses to open-ended questions marked "test").

Responses that were not coherent were labeled, but not excluded (study purpose: *n* = 5; previously: participated *n* = 3).

Procedure

After providing informed consent, participants read and answered questions about three vignettes that described counterfeit-object cases. Each participant responded to three condition and vignette combinations randomly assigned on each factor without replacement such that all participants saw each vignette (Darrel, Emma, Gerald) and each condition (Ignorance, Knowledge, Gettier) exactly once. After reading each vignette, participants responded to a series of items in a fixed order on separate screens. Items were presented as follows: knowledge attribution, comprehension check, reasonableness judgment, luck attribution (two items), and alternative knowledge probe. Next, participants answered questions related to their experience completing the study, data exclusion criteria, and demographics, respectively. Finally, participants were debriefed and compensated if applicable.

Analytic Approach

Analyses were conducted on combined raw data collected in SocSciSurvey and Qualtrics. In the original protocol, we planned to evaluate the quality of each student-led team's data, including the raw data, analysis scripts, codebooks, cleaned data sets, and narrative summaries of results. We also planned that data would be included in analyses only if teams received a CREP completion certificate after these products passed a quality check. However, the original protocol did not describe clear criteria that would be used to detect and correct errors, and many teams did not submit their projects for final CREP review. In order to conduct reproducible, transparent analyses, we chose not to exclude data from teams who failed to meet the target sample size or did not receive completion certificates. All teams were required to receive CREP approval before commencing data collection; this process included preparing an OSF page with all materials and videos of their procedure, submitting the page for review by CREP reviewers, and making any revisions as necessary. If data collection teams received approval and collected their data using the centralized survey, their data was also included in analysis. Because of this oversight and the strict data quality exclusions implemented at the level of participants, we were not concerned about team level variation in data quality. Still, we repeated our primary analyses excluding data from the teams that did not receive completion certificates. Generally, we observed the same patterns of results (see https://osf.io/nvfbm).[16] A summary of how the teams independently analyzed their data (i.e., the test used for the effect of condition on knowledge attribution) is reported in the last column of Table 1, and those results can be found on their OSF pages.

Multilevel Models

Multilevel models were used to evaluate our hypotheses. The unit of analysis was the question response, and cross-classified random intercepts for the vignette, participant, and data collection site were included to account for the nesting of responses within these groups.[17] Exact model specification can be found at https://osf.io/8ut6e/.

Assumptions and Transformations. While the approved protocol described testing assumptions before conducting analyses, it did not detail criteria that would be used for testing assumptions or approaches to handling model convergence issues. No convergence issues

[16] Analyses were repeated using the original exclusion criteria, which included 5 additional participants who reported ages 100 or above and excluded participants from sites without CREP completion certifications. One minor difference in results was found. For the “reasonableness” dependent variable, the vignette by condition interaction was not observed in one of the tested models, likely because of the smaller sample size after exclusions.

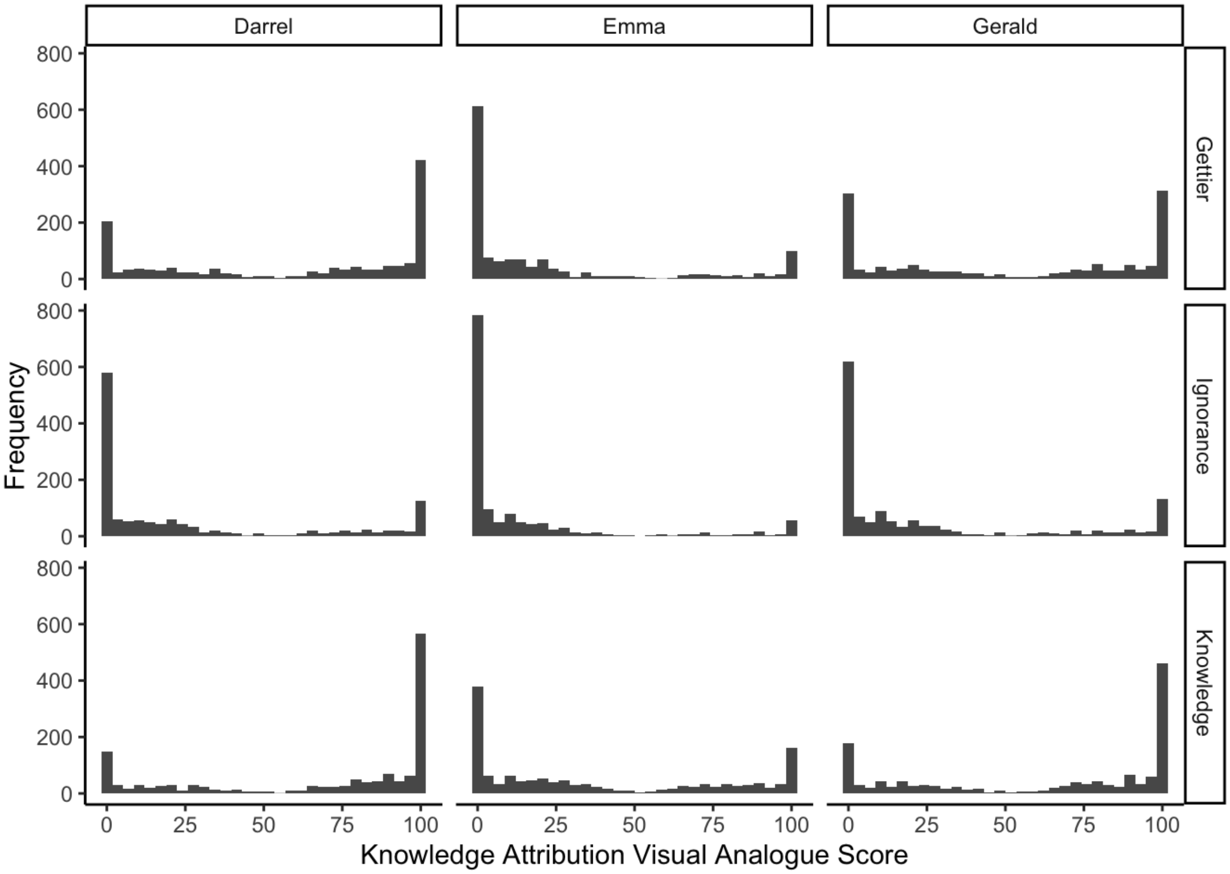
[17] In the approved protocol, data collection was described as taking place in labs. Labs were described as uniquely identifying data collection sites. However, at some data collection sites, multiple student-led teams joined this project (e.g., under the mentorship of the same PI, multiple students joined the project as “labs”). Observations were labeled as belonging to both a “lab” (which we describe as a “student-led team”) and a data collection site. For analyses, the data collection site was used in place of the “lab” variable described in the approved protocol.

emerged during analyses. Here we describe the approach taken to test assumptions. Assumptions of and related to linearity are primarily relevant for the analysis of the continuously measured dependent variables. The continuous knowledge attribution variable was bimodal overall and within vignette and condition combinations (see Figure 2).

To examine normality, homogeneity, and linearity, we used linear mixed models that predicted continuously measured knowledge, reasonableness, and luck attribution as a function of condition with covariates of compensation, age, gender, and education. The residual distributions were also bimodal or heavily skewed, indicating violations of the residual normality assumption. Further, plots of residuals by fitted values suggested that residuals varied as a function of predicted values, indicating violations of the homoscedasticity assumption. Last, and most importantly, the linearity assumption was not met for any dependent variable which each showed a sigmoid function similar to binary outcome data.

Figure 2

Knowledge Attribution Visual Analogue Scores by Vignette and Condition



Transforming continuous variables into discrete variables for analysis is not generally recommended (MacCallum et al., 2002; Maxwell & Delaney, 1993). For the present analyses, however, this approach was necessary due to the already bimodal distribution of the dependent variables and the suggested sigmoid function from the residual data screening results. Thus, we split the continuously measured versions of the three dependent variables such that scores at and below 40 and scores at and above 60 were classified into discrete categories. Higher scores were coded as 1 to indicate knowledge, reasonableness, or ability, and lower scores were coded as 0 to indicate belief, unreasonableness, or luck. We chose these points so that participants clearly had indicated a side (i.e., 41-59 were considered neutral), and very few data points were lost in this middle range. Of the non-missing responses on each continuous measure, 359 (2.87%) responses were dropped for the knowledge attribution variable, 279 (2.23%) responses were dropped for the reasonableness attribution variable, and 683 (5.85%) responses were dropped for the luck attribution variable.

This approach allowed us to validly interpret model results and also test whether the method of measurement (continuous or binary) affected results. Data screening was examined for logistic models with the same parameters as above; the assumptions of logistic regression were met: no empty or small categories, linearity of the logit for continuous predictors, and additivity of the predictors. We repeated our primary analyses with the continuous dependent measures using linear regressions to see whether this deviation impacted our findings. Overall, we found the same pattern of results.[18] See https://osf.io/nvfbm for details.

Model Steps. A series of multilevel logistic regression models were fit predicting knowledge attributions and reasonableness judgments. Transformed and originally binary responses were analyzed together. Each model was fit including all participants with no missing data on that model's variables. After estimating a baseline intercept-only model (Model 1), we fit models with random intercepts for vignette (Model 2), person (Model 3), and data collection site (Model 4) added sequentially. In Model 5, participant age, compensation, gender, and education (in years) were added as fixed effects. These variables served as covariates and were included in our original analysis plan due to previous research that demonstrated their impact on knowledge attribution. Finally, the knowledge condition variable was added in Model 6. To see if the effect of condition varied by vignette, the interaction between vignette and condition was added as a

[18] The only difference we found in comparing results of the linear versus logistic models was in the sample source analyses for the reasonableness and knowledge dependent measures. The linear models found interaction effects between condition and sample source (MTurk vs. not MTurk) where the logistic models did not. Examination of the patterns of results indicated the same condition differences for both data sources with slightly weaker effects for the MTurk data than the non-MTurk data.

fixed effect in Model 6A. Additional models were fit to test the moderating effects of participant source (Model 6B; MTurk vs. lab), luck attributions (Model 6C; luck vs. ability), and the original measurement scale (Model 6D; binary vs. continuous). The conceptual models presented in 1-6B were preregistered, maintaining independent and random effects variables in the updated analysis plan. Model 6D was added when the data screening indicated the VAS results were not continuous as expected and the dependent variables were dichotomized. The exact implementation of the multilevel models (i.e., model order and interpretation) were updated from our preregistered plan to ensure appropriate statistical inference (see Appendix A for full details).

Results

To better test our research questions, we implemented analyses that differed from those we originally planned.[19] All deviations are summarized in Appendix A. The results section as it appeared in the approved protocol is also included in Appendix A with updated statistics where possible. While the results below indicate components of the random structure (i.e., intercepts of participant and site) do not add to or improve the models, we included these facets to match the preregistered plan and to maintain independence of observations (i.e., participant intercepts are arguably necessary for a repeated measures design). The lack of participant variance suggests that individuals did not systematically vary in their responding across vignette-condition combinations; the lack of site variance suggests that results were consistent across data collection sites.

[19] In the approved protocol, the results section focused heavily on the project's logistics and structured results reporting in ways that would not allow for a transparent and thorough description of model fit and other important aspects of results, like assumption checks. Further, some model specification details in the approved protocol conflicted with stated research questions (e.g., we specified that the null model would include the focal predictor, which would have rendered the null model invalid as null models are not supposed to include any predictors).

For each focal model, we report the model fit statistics and parameter estimates. Parameter estimates for logistic models can be interpreted in a similar fashion to linear regression models: negative values indicate that increasing the predictor decreases the likelihood of the dependent variable (e.g., the choice coded as one, therefore, increasing the likelihood of the choice coded as zero), and positive values indicate that increases in the predictor correspond to increases in the likelihood of the dependent variable (e.g., the choice coded as one). When predictors are also categorical, increasing the predictor indicates a comparison between the predictor group coded as zero and the predictor group coded as one. All pseudo-R^2 values were calculated with the *MuMIn* package (Bartoń, 2020) using formulas for fixed and random effects from Nakagawa et al. (2017).

Knowledge Attribution

The goal of the present research was to provide a well-powered estimate of the magnitude and prevalence of Gettier intuitions (i.e., the difference in knowledge attribution between Gettier and knowledge conditions) across different vignettes and testing sites in a replication and extension of Turri et al. (2015). Models were fit in steps to determine whether participants attributed knowledge to the protagonist at different rates as a function of condition. See Table 5 for a summary of model results. Compared to the baseline Model 1 (AIC = 18881.09), the model including random intercepts for vignette (AIC = 17834.75) explained more variance (Pseudo-R^2 = .08 - .10). Participants attributed knowledge most frequently in response to the Darrel vignette (52.16%) and least frequently in response to the Emma vignette (20.94%). See Table 6 for differences by vignette extracted from Model 2.

Table 5

Knowledge Attribution Model Summaries

Parameter Estimate or	**Model**						
Statistic	**1**	**2**	**3**	**4**	**5**	**6**	**6A**
Fixed Effects							
Intercept	-0.44 (0.02)***	-0.49 (0.35)	-0.49 (0.35)	-0.49 (0.35)	-0.37 (0.36)	-0.18 (0.40)	0.56 (0.13)***
Age					0.003 (0.00)	0.004 (0.00)	0.004 (0.00)
Gender					-0.07 (0.04)	-0.08 (0.04)*	-0.09 (0.04)*
Education					-0.02 (0.01)*	-0.02 (0.01)*	-0.02 (0.01)*
Compensation					0.02 (0.04)	0.02 (0.04)	0.02 (0.04)
Condition: Ignorance						-1.31 (0.05)***	-1.60 (0.08)***
Condition: Knowledge						0.61 (0.04)***	0.50 (0.08)***
Vignette: Emma							-1.93 (0.08)***
Vignette: Gerald							-0.40 (0.07)***
Ignorance X Emma							0.98 (0.13)***
Ignorance X Gerald							0.21 (0.11)*
Knowledge X Emma							0.40 (0.11)***
Knowledge X Gerald							0.02 (0.11)
Random Effects							
Site				< 0.001	< 0.001	< 0.001	< 0.001
Participant			< 0.001	< 0.001	< 0.001	< 0.001	< 0.001
Vignette		0.600	0.600	0.600	0.601	0.669	< 0.001
AIC	18881.09	17834.75	17836.75	17838.75	17554.31	15871.99	15807.69

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest increasing likelihood of knowledge attribution. For condition, the comparison group was Gettier, and for vignette, the comparison group was Darrel. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.

* *p* < .05; ** *p* < .01; *** *p* < .001

Table 6

Knowledge Attributions from Model 2 Overall and by Vignette

	Overall	Darrel	Emma	Gerald
Believes	8595 (60.92%)	2268 (48.12%)	3716 (78.95%)	2611 (55.70%)
Knows	5513 (39.08%)	2445 (51.88%)	991 (21.05%)	2077 (44.30%)

The model nesting vignette within participants (Model 3; AIC = 17836.75) explained similar amounts of variance (Pseudo-R^2 = .08 - .10) as Model 2. The addition of the random effect of data collection site in Model 4 (AIC = 17838.75) likewise did not improve model fit (Pseudo-R^2 = .08 - .10). The model including the covariates predicting knowledge attributions as fixed effects (Model 5; AIC = 17554.31) was more useful in explaining variance in knowledge attribution than previous models. Age predicted knowledge attribution, such that as age increased, participants were more likely to attribute knowledge to the protagonists. Education was a negative predictor; rates of knowledge attribution decreased as reported education increased. However, these fixed effects accounted for a very small proportion of the variance, Pseudo-R^2 < .001.

Model 6 served as the key replication test of Turri et al. (2015). The knowledge condition was added as a fixed effect (AIC = 15539.57). This model performed better than the previous model and revealed an effect of condition on knowledge attribution (Pseudo-R^2 = .12 - .15). See Table 5 for model statistics and Table 7 for knowledge attribution rates by condition. Participants were more likely to attribute knowledge to the protagonist in the knowledge condition vignette than to the protagonists in the ignorance and Gettier condition vignettes; further, the ignorance condition differed from the Gettier condition. Thus, we did not fully replicate the results of Turri et al., who found no difference in knowledge attribution between the knowledge and Gettier conditions. Using the data from this model, each condition was examined for difference from chance using χ^2 tests. In the knowledge condition, participants were more likely than chance to attribute knowledge to the protagonist. Participants were less likely than chance to attribute knowledge to the protagonists in the Ignorance and Gettier condition vignettes, all *p*s < .001 (see Table 7).

Table 7

Knowledge Attributions from Model 6 Overall and by Condition

	Overall	Knowledge	Ignorance	Gettier
Believes	8476 (61.00%)	2005 (43.41%)	3833 (82.06%)	2638 (57.29%)
Knows	5419 (39.00%)	2614 (56.59%)	838 (17.94%)	1967 (42.71%)
$\chi^2(1)$		80.29***	1920.37***	97.77***
Darrel				
Believes	2239 (48.24%)	454 (28.73%)	1170 (76.82%)	615 (39.99%)
Knows	2402 (51.76%)	1126 (71.27%)	353 (23.18%)	923 (60.01%)
Gerald				
Believes	2570 (55.68%)	558 (36.83%)	1255 (79.63%)	757 (49.64%)
Knows	2046 (44.32%)	957 (63.17%)	321 (20.37%)	768 (50.36%)
Emma				
Believes	3667 (79.06%)	993 (65.16%)	1408 (89.57%)	1266 (82.10%)
Knows	971 (20.94%)	531 (34.84%)	164 (10.43%)	276 (17.90%)

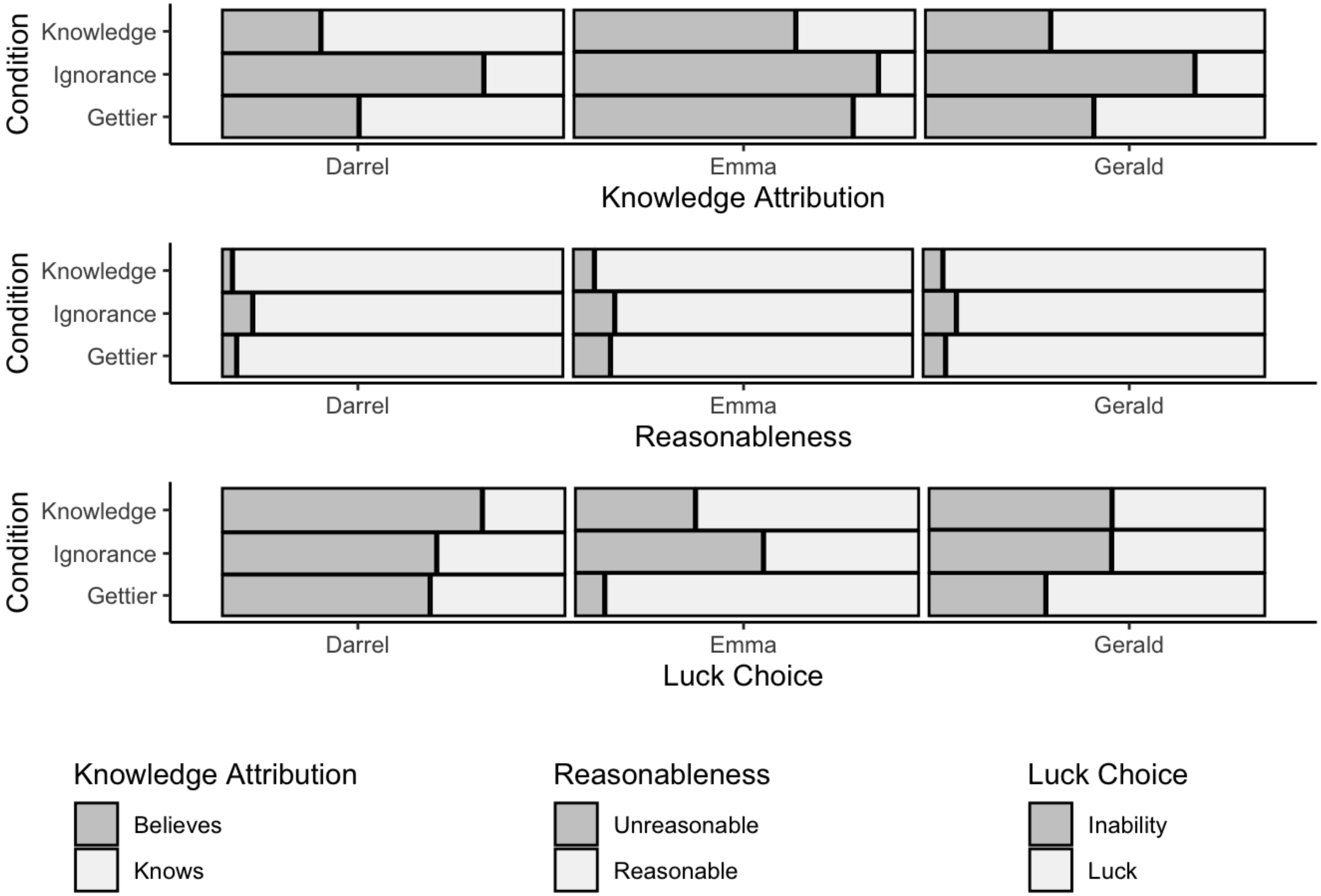
Note. χ^2 tests comparing participant knowledge attributions in each condition to chance were conducted using data from Model 6.
***$p < .001$

Does the effect of condition on knowledge attributions differ by vignette?

To better understand whether the effect of condition varied as a function of the vignette's content, Model 6A was estimated including an interaction between vignette and condition (AIC = 15807.69). This model fit the data better (Pseudo-R^2 = .20 - .24) than Model 6. As shown in Figure 3, the pattern of results was the same for every vignette; however, values suggest that the interaction between condition and vignette accounted for some of the variance in knowledge attributions. The size of the differences between conditions (and between vignettes) depended on the vignette-condition combinations.

Figure 3

Knowledge attribution, reasonableness, and luck/(in)ability rates by vignette and condition.



In responding to the Darrel vignette, participants attributed knowledge at different rates according to the vignette's condition, $\chi^2(2) = 781.00$, $p < .001$. Participants were more likely to attribute knowledge when responding to the Gettier condition version ($\hat{p} = .60$) than in the ignorance condition version ($\hat{p} = .23$; Cramér's $V = .37$, 95% CI [.34, .41], $\chi^2(1) = 425.61$, $p < .001$). They were also more likely to attribute knowledge to Darrel when responding to the knowledge condition version ($\hat{p} = .71$) than in the Gettier condition version ($\hat{p} = .60$; Cramér's $V = .12$, 95% CI [.08, .15], $\chi^2(1) = 43.30$, $p < .001$).

The pattern of responding was similar for the Emma vignette; the likelihood that participants attributed knowledge to Emma differed according to the vignette's condition, $\chi^2(2) = 291.42$, $p < .001$. Participants were more likely to attribute knowledge when responding to the Gettier condition of the Emma vignette ($\hat{p} = .18$) than in the ignorance condition of the Emma vignette ($\hat{p} = .10$; Cramér's $V = .11$, 95% CI [.07, .14], $\chi^2(1) = 35.15$, $p < .001$). The likelihood of knowledge attribution was higher for the knowledge version of the vignette ($\hat{p} = .35$) than for the Gettier version ($\hat{p} = .18$; Cramér's $V = .19$, 95% CI [.16, .23], $\chi^2(1) = 112.59$, $p < .001$).

In response to the Gerald vignette, participant knowledge attributions similarly differed according to vignette condition, $\chi^2(2) = 607.03$, $p < .001$. Participants were more likely to attribute knowledge in response to the Gettier condition version of the Gerald vignette ($\hat{p} = .50$) than to the ignorance condition version of the Gerald vignette ($\hat{p} = .20$; Cramér's $V = .31$, 95% CI [.28, .35], $\chi^2(1) = 304.67$, $p < .001$). In addition, they were more likely to attribute knowledge to Gerald in the knowledge condition version ($\hat{p} = .63$) than in the Gettier condition version ($\hat{p} = .50$; Cramér's $V = .13$, 95% CI [.09, .17], $\chi^2(1) = 50.27$, $p < .001$).

To interpret the condition by vignette interaction, we examined Cramér's *V* for the analyses of each vignette. This approach revealed that the likelihood of knowledge attributions in the Gettier and ignorance conditions differed less for the Emma vignette than for the Darrel and Gerald vignettes. Additionally, the Gettier and knowledge conditions of the Darrel vignette produced a smaller difference in likelihood than that for those conditions of the other two vignettes. Thus, participants demonstrated Gettier intuitions in all three vignettes (i.e., participants were more likely to deny knowledge in the Gettier condition than in the knowledge condition, a case of justified true belief), but these Gettier intuitions were weakest in response to the Darrel vignette and strongest in response to the Emma vignette.

Reasonableness Judgments

As a secondary dependent measure, judgments of reasonableness were predicted in a series of logistic regression models paralleling those for knowledge attributions. See Table 8 for a summary of model results. Compared to a baseline intercept-only model (Model 1, AIC = 7343.35), a model with a random intercept for vignette (Model 2, AIC = 7286.55) explained more variance. The likelihood of the protagonist being judged as reasonable varied by vignette (Pseudo-R^2 = .00 - .02); although, overall participants were far more likely to respond that the protagonist was reasonable than unreasonable in all three vignettes. Collapsing across conditions, participants were more likely to judge Emma as unreasonable than Gerald. Participants were more likely to judge Gerald as unreasonable than Darrel (see Table 9).

Table 8

Reasonableness Judgment Model Summaries

Parameter Estimate or	**Model**						
Statistic	**1**	**2**	**3**	**4**	**5**	**6**	**6A**
Fixed Effects							
Intercept	2.56 (0.03)***	2.59 (0.16)***	2.59 (0.16)***	2.59 (0.16)***	1.81 (0.24)***	1.84 (0.25)***	2.43 (0.22)***
Age					-0.005 (0.00)	-0.005 (0.00)	-0.005 (0.00)
Gender					-0.18 (0.07)*	-0.18 (0.07)*	-0.18 (0.07)*
Education					0.06 (0.01)***	0.06 (0.01)***	0.06 (0.01)***
Compensation					0.25 (0.07)**	0.25 (0.07)***	0.25 (0.07)***
Condition: Ignorance						-0.40 (0.08)***	-0.86 (0.16)***
Condition: Knowledge						0.43 (0.09)***	0.39 (0.20)
Vignette: Emma							-1.10 (0.16)***
Vignette: Gerald							-0.52 (0.17)**
Ignorance X Emma							0.74 (0.20)***
Ignorance X Gerald							0.42 (0.21)*
Knowledge X Emma							0.26 (0.25)
Knowledge X Gerald							-0.24 (0.26)
Random Effects							
Site				< 0.001	< 0.001	0.023	< 0.001
Participant			< 0.001	< 0.001	< 0.001	0.091	0.046
Vignette		0.274	0.274	0.274	0.278	0.277	< 0.001
AIC	7343.35	7286.55	7288.56	7290.55	7144.10	7047.13	7025.80

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest increasing likelihood of reasonableness judgments. For condition, the comparison group was Gettier, and for vignette, the comparison group was Darrel. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.

* $p < .05$; ** $p < .01$; *** $p < .001$

Table 9

Reasonableness Judgments from Model 2 Overall and by Vignette

	Overall	Darrel	Emma	Gerald
Unreasonable	1021 (7.19%)	237 (5.01%)	447 (9.48%)	337 (7.10%)
Reasonable	13173 (92.81%)	4493 (94.99%)	4269 (90.52%)	4411 (92.90%)

A model with a random intercept for vignette nested within participant (Model 3, AIC = 7288.56) explained similar amounts of variance (Pseudo-R^2 = .00 - .02) as Model 2 . The model with a random intercept for vignette nested in participant nested in data collection site (Model 4, AIC = 7290.55) did not explain more variance (Pseudo-R^2 = .00 - .02) than previous models. In Model 5, covariates were added as fixed effects (AIC = 7144.10). Relative to Model 4, this model was more useful in explaining variance in judgments of reasonableness (Pseudo-R^2 = .01 - .04). Participant compensation, gender, and education were associated with reasonableness judgments. Participants who were compensated and female participants were more likely to judge the protagonist as reasonable than uncompensated and male participants. As the participant's years of education increased, the likelihood that they would judge the protagonist as reasonable increased.

Finally, we estimated a model including knowledge condition as a fixed effect (Model 6, AIC = 7047.13). This model performed better than Model 5 and revealed an effect of condition on reasonableness judgment (Pseudo-R^2 = .01 - .05). Participants were more likely to judge the protagonist in the knowledge condition vignette as reasonable than the protagonists in the other two conditions (see Table 10). Protagonists in the ignorance condition vignette were less likely to be judged as reasonable than protagonists in the knowledge and Gettier condition vignettes.

Table 10

Reasonableness Judgments from Model 6 Overall and by Condition

	Overall	Knowledge	Ignorance	Gettier
Unreasonable	1007 (7.21%)	217 (4.65%)	467 (10.02%)	323 (6.94%)
Reasonable	12967 (92.79%)	4447 (95.35%)	4192 (89.98%)	4328 (93.06%)

Does the effect of condition on reasonableness judgments differ by vignette?

To test whether the effect of condition on reasonableness judgments varied by vignette, a model was estimated that included an interaction between vignette and condition (Model 6A, AIC = 7025.80). This model explained more variance than the model without the interaction term. As shown in Figure 3, although the general pattern was the same for all vignettes, the magnitudes of the differences varied by vignette (Pseudo-R^2 = .02 - .08).

The likelihood that participants judged the protagonist as reasonable varied by condition in response to the Darrel vignette, $\chi^2(2) = 781.00$, $p < .001$, Emma vignette $\chi^2(2) = 36.36$, $p < .001$, and Gerald vignette $\chi^2(2) = 21.10$, $p < .001$. Participants were more likely to judge Darrel to be reasonable in the Gettier condition vignette ($\hat{p}$ = .96) than in the ignorance condition vignette ($\hat{p}$ = .91; Cramér's V = .10, 95% CI [.06, .13], $\chi^2(1) = 28.84$, $p < .001$), but we found no evidence that reasonableness judgments differed between participants responding to the Gettier and knowledge conditions of that vignette (Cramér's V = .03, 95% CI [.02, .07], $\chi^2(1) = 3.44$, $p = .064$). The same pattern of results appeared in response to the Gerald vignette; participants were more likely to judge Gerald as reasonable when responding to the Gettier condition vignette ($\hat{p}$ = .94) as opposed to the ignorance condition vignette ($\hat{p}$ = .91; Cramér's V = .06, 95% CI [.03, .09], $\chi^2(1) = 10.49$, $p = .001$), but the Gettier and knowledge vignettes produced similar rates of reasonableness judgments, ($\hat{p}$ = .94; Cramér's V = .02, 95% CI [.02, .05], $\chi^2(1) = 0.77$, $p = .381$).

The condition by vignette interaction in predicting judgments of reasonableness appears to have emerged because of the condition differences produced by the Emma vignette. While participants were equally likely to judge Emma as reasonable in the Gettier and ignorance conditions, (Cramér's V = .02, 95% CI [.02, .06], $\chi^2(1) = 1.12$, $p = .291$), participants were more likely to judge Emma as reasonable in response to the knowledge condition vignette ($\hat{p}$ = .94) than in response to the Gettier condition vignette ($\hat{p}$ = .89; Cramér's V = .09, 95% CI [.05, .12], $\chi^2(1) = 22.44$, $p < .001$). Thus, condition differences were found between the Gettier and ignorance versions of the Darrel and Gerald vignettes, but not the Emma vignette, and between the Gettier and knowledge versions of the Emma vignette, but not the Darrel and Gerald vignettes.

Participant Recruitment

Data were collected from MTurk workers as well as participants recruited from individual labs. As the MTurk sample more likely represented the sample originally collected by Turri et al., we examined whether participant recruitment moderated the effect of condition on knowledge attributions and reasonableness judgments. Though Model 6B (AIC = 15850.16) was superior to Model 6, the interaction term was not a significant predictor of knowledge attributions (ΔPseudo-R^2 = .00 - .01). Next, we estimated the same model (Model 6B) in predicting judgments of reasonableness (AIC = 7017.37). While this model performed better than Model 6, the interaction between condition and recruitment type was not significant (ΔPseudo-R^2 = .00 - .01). See Table 11 for summary of results.

Table 11

Participant Recruitment Moderation Model (6B) Summaries

	Measure	
Parameter or Statistic	**Knowledge**	**Reasonableness**
Fixed Effects		
Intercept	-0.07 (0.41)	2.07 (0.25)***
Age	< 0.001 (0.00)	-0.01 (0.003)***
Gender	-0.11 (0.04)**	-0.23 (0.07)**
Education	-0.02 (0.01)*	0.06 (0.01)***
Compensation	-0.03 (0.05)	0.13 (0.08)
Condition: Ignorance	-1.29 (0.05)***	-0.38 (0.08)***
Condition: Knowledge	0.59 (0.05)***	0.44 (0.09)***
Participant Source	0.32 (0.13)*	1.39 (0.37)***
Source X Ignorance	-0.33 (0.19)	-0.66 (0.43)
Source X Knowledge	0.30 (0.17)	-0.44 (0.51)
Random Effects		
Site	< 0.001	0.049
Participant	< 0.001	< 0.001
Vignette	0.670	0.278
AIC	15850.16	7017.37

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest increased likelihood of knowledge attributions or reasonableness judgments. Source was coded with lab participants as the comparison group. For condition, the comparison group was Gettier. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.
* *p* < .05; ** *p* < .01; *** *p* < .001

Exploratory Analyses

In addition to the hypotheses and research questions outlined in the approved protocol, we conducted additional exploratory analyses to examine three additional research questions and assess the influence of original measurement characteristics (binary vs. continuous).

"Direct" Replication Analysis

As previously explained, the design of our study substantially differed from that of Turri et al. (2015, Experiment 1). Rather than encountering one of three conditions of the Darrel/Squirrel vignette, our participants viewed three conditions matched with three vignettes in a within-participants design. Perhaps our observation of a Gettier intuition effect, which was not found in the original experiment, can be explained by these methodological changes. To explore this possibility, we compared the knowledge attribution rates of participants who viewed the Darrel vignette first (*n* = 2538) in an analysis devised to closely approximate Turri et al.'s original test.[20] Overall, participants attributed knowledge at different rates according to condition $\chi^2(2) = 252.57$, $p < .001$, Cramér's $V = .34$, 95% CI [.30, .38], and the pattern of effects mirrored those of our primary analysis. Participants responding to the Gettier condition were more likely to attribute knowledge to Darrel ($\hat{p} = .59$) than those responding to the ignorance condition ($\hat{p} = .32$), $\chi^2(1) = 103.61$, $p < .001$, Cramér's $V = .26$, 95% CI [.22, .32]. However, participants were less likely to attribute knowledge in response to the Gettier condition vignette than to the knowledge condition vignette ($\hat{p} = 0.72$), $\chi^2(1) = 30.48$, $p < .001$, Cramér's $V = .14$, 95% CI [.10, .20]. Thus, this analysis provided further evidence for Gettier intuitions despite more closely approximating Turri's original test than our planned analysis. These effects were similar for the Gerald vignette when presented as the first vignette (i.e., same effect size and pattern) and the Emma vignette (i.e., same pattern and half the effect size).

Luck Attributions

Attributions of luck were predicted in a series of multilevel logistic regressions models. These models were fit in the same fashion as the models focused on the two dependent variables, with one notable difference: observations where the participant did not correctly answer the first part of our two-part luck attribution measure were excluded. That is, the luck versus ability attributions that followed incorrect identification responses were excluded from analyses (*n* =

[20] Only participants who missed the Darrel comprehension check question (*n* = 1138) were excluded from this analysis to replicate the exclusion criteria implemented in the original experiment.

952; 6.58%). See Table 12 for summary of Models 1-6A. Compared to the baseline intercept-only model (Model 1, AIC = 11269.61), a model with a random intercept for vignette (Model 2, AIC = 10613.78) explained more variance. The likelihood that outcomes were attributed to luck varied according to vignette (Pseudo-R^2 = .08 - .09). While the Darrel vignette produced more attributions to ability than luck, the Emma vignette produced more attributions to luck than ability (see Table 13).

Table 12

Luck/(in)ability Attribution Model Summaries

Parameter Estimate or Statistic (SE)	**Model**						
	1	**2**	**3**	**4**	**5**	**6**	**6A**
Fixed Effects							
Intercept	-0.04 (0.02)*	-0.04 (0.34)	-0.04 (0.34)	-0.04 (0.34)	0.40 (0.36)	-0.25 (0.37)	0.91 (0.13)***
Age					0.004 (0.00)	0.004 (0.00)*	0.004 (0.00)*
Gender					0.01 (0.04)	0.01 (0.04)	-0.004 (0.04)
Education					-0.03 (0.01)***	-0.03 (0.01)***	-0.03 (0.01)***
Compensation					-0.17 (0.04)***	-0.18 (0.04)*	-0.20 (0.04)***
Condition: Ignorance						1.03 (0.05)***	0.08 (0.08)
Condition: Knowledge						0.94 (0.05)***	0.72 (0.08)***
Vignette: Emma							-2.85 (0.11)***
Vignette: Gerald							-1.07 (0.08)***
Ignorance X Emma							2.53 (0.14)***
Ignorance X Gerald							0.72 (0.11)***
Knowledge X Emma							1.07 (0.14)***
Knowledge X Gerald							0.10 (0.11)
Random Effects							
Site				< 0.001	< 0.001	0.099	0.022
Participant			< 0.001	< 0.001	< 0.001	0.066	0.040
Vignette		0.581	0.581	0.581	0.585	0.611	< 0.001
AIC	17776.67	16771.30	16773.30	16775.30	16489.60	15896.17	15458.37

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest increasing likelihood of ability attributions. For condition, the comparison group was Gettier, and for vignette, the comparison group was Darrel. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.

* *p* < .05; ** *p* < .01; *** *p* < .001

Table 13

Luck (In)ability Attributions from Model 2 Overall and by Vignette

	Overall	Darrel	Emma	Gerald
Luck	6551 (51.08%)	1434 (33.36%)	2910 (67.50%)	2207 (52.34%)
(In)ability	6275 (48.92%)	2864 (66.64%)	1401 (32.50%)	2010 (47.66%)

A model with a random intercept for vignette nested within participants (Model 3, AIC = 16773.30) explained similar amounts of variance as the previous model (Pseudo-R^2 = .08 - .09). Nesting within the data collection site ((Model 4, AIC = 16775.30) did not improve the model fit (Pseudo-R^2 = .08 - .09). Next, covariates were added to the model as fixed effects (Model 5, AIC = 16489.60). Relative to Model 4, Model 5 explained more variance in luck attributions (Pseudo-R^2 = .08 - .10). Years of education, age, and compensation independently predicted luck attributions (see Table 12).

Finally, we estimated a model including condition as a fixed effect (Model 6, AIC = 15896.17). This model performed better than the previous models; the likelihood of luck attributions differed according to condition (Pseudo-R^2 = .05 - .06). Participants were more likely to attribute the outcome to luck in the Gettier condition than in the other two conditions (see Table 14). In response to both the knowledge condition and the ignorance condition, participants were more likely to attribute outcomes to the protagonist's ability than to luck, but they were more likely to make luck attributions than ability attributions in response to the Gettier condition vignette.

Table 14

Luck/(In)ability Attributions from Model 6 Overall and by Condition

	Overall	Knowledge	Ignorance	Gettier
Luck	6451 (51.04%)	1888 (44.53%)	1784 (42.71%)	2779 (65.81%)
(In)ability	6189 (48.96%)	2352 (55.47%)	2393 (57.29%)	1444 (34.19%)

Vignette interactions. To better understand whether the effect of condition on luck attributions varied as a function of vignette, we estimated a model including an interaction between vignette and condition (Model 6A, AIC = 15458.37). This model explained more variance (Pseudo-R^2 = .20 - .23) than Model 6. As shown in Figure 3, each vignette demonstrated a different pattern of effects. Post hoc analyses suggested that the vignette by condition interaction was driven by responses to the Gettier condition. The difference in likelihoods of luck attributions between the Gettier and ignorance conditions was absent for the Darrel vignette (Cramér's V = .02, p = .315), moderate for the Gerald vignette (Cramér's V = .20, p < .001), and large for the Emma vignette (Cramér's V = .50, p < .001). The difference in luck attributions between the Gettier and knowledge conditions was largest in responses to the Emma vignette (Cramér's V = .32, p < .001) but of similar size in response to the Darrel vignette (Cramér's V = .16, p < .001) and Gerald vignette (Cramér's V = .20, p < .001).

Luck/(in)ability as a moderator. Next, we explored whether attributions of outcomes to luck versus ability influence knowledge attributions as suggested by prior research (Turri, 2016, 2017). Turri (2016, Experiment 7) found a strong positive correlation between knowledge attributions and attributions to ability rather than luck (r = .622) and a moderating effect of luck attributions on Gettier intuitions; participants attributed knowledge less frequently when protagonists were perceived as having arrived at a truth because of a lucky guess rather than because of their ability ($\eta_p^2 = .353$; Turri, 2016, Experiment 7).

We tested whether luck attributions moderated the effect of condition on knowledge attribution among participants who accurately identified that the protagonist was correct (in the Gettier and knowledge conditions) or incorrect (in the ignorance conditions) in their identification of the object as real or counterfeit. The main effect of luck attributions and the interaction between condition and luck attributions were added to Model 6 of the knowledge attributions analysis (Model 6C; AIC = 13363.98). This model (Pseudo-R^2 =.24 - .28) explained more variance in knowledge attributions than Model 6. See Table 15 for model summary.

Table 15

Luck/(In)ability Attribution Moderation Model (6C) Summary

Parameter or Statistic	**Knowledge**
Fixed Effects	
Intercept	-0.48 (0.36)
Age	0.003 (0.00)
Gender	-0.13 (0.05)**
Education	-0.02 (0.01)*
Compensation	0.02 (0.05)
Condition: Ignorance	-1.00 (0.08)***
Condition: Knowledge	0.35 (0.66)***
Luck/Ability	1.03 (0.07)***
Ignorance X Luck/Ability	-1.13 (0.11)***
Knowledge X Luck/Ability	0.20 (0.10)
Random Effects	
Site	< 0.001
Participant	< 0.001
Vignette	0.574
AIC	13363.98

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest increased likelihood of knowledge attributions. Luck/(in)ability was coded so that 0 indicated luck and 1 indicated (in)ability. For condition, the

comparison group was Gettier. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.
* $p < .05$; ** $p < .01$; *** $p < .001$

Condition affected knowledge attributions when participants attributed the protagonists' (in)correct identification to bad or good luck, $\chi^2(2) = 211.03$, $p < .001$. Participants were more likely to attribute knowledge to the protagonist in the Gettier condition vignette ($\hat{p} = .31$) than in the ignorance condition vignette ($\hat{p} = .17$; Cramér's $V = .16$, 95% CI [.13, .19], $\chi^2(1) = 116.03$, $p < .001$). They were also more likely to attribute knowledge in the knowledge condition vignette ($\hat{p} = .38$) than in the Gettier condition vignette ($\hat{p} = .31$; Cramér's $V = .07$, 95% CI [.05, .10], $\chi^2(1) = 24.54$, $p < .001$).

Similarly, condition affected knowledge attributions when participants attributed the protagonists' (in)correct identification to (in)ability $\chi^2(2) = 1737.19$, $p < .001$. Participants in this group were more likely to attribute knowledge to the protagonist in the Gettier condition vignette ($\hat{p} = .66$) than in the ignorance condition vignette ($\hat{p} = .16$; Cramér's $V = .51$, 95% CI [.48, .54], $\chi^2(1) = 972.07$, $p < .001$). These participants were also more likely to attribute knowledge in the knowledge condition vignette ($\hat{p} = .73$) than in the Gettier condition vignette ($\hat{p} = .66$; Cramér's $V = .08$, 95% CI [.05, .11], $\chi^2(1) = 21.36$, $p < .001$). While the knowledge attribution difference between the Gettier and ignorance conditions was larger when participants made ability attributions (Cramér's $V = .51$) than when they made luck attributions (Cramér's $V = .16$), effect sizes were similar for the differences between the Gettier and knowledge conditions (Cramér's $V = .08$ vs. .07). Thus, unlike in previous research (Turri, 2016, 2017), luck attributions did not decrease the likelihood of participants demonstrating Gettier intuitions.

Alternative Knowledge Probe

We also assessed whether question wording affected participants' knowledge attributions as has been suggested by previous research (e.g., Machery et al., 2017b; Nagel, San Juan, et al., 2013). Participants may be more likely to deny knowledge to a protagonist when they are asked a more nuanced question (whether the protagonist knew or only felt like they knew but did not actually know; Nagel, San Juan, et al., 2013) than when they are asked a simpler question (whether the protagonist knew or did not know).

In our exploratory analyses of the alternative knowledge probe (i.e., following Model steps 1 through 6), we found a pattern of results similar to those for the analyses of our primary knowledge measure (Model 6: AIC = 16332.68; Pseudo-R^2 = .16 - .21). See Table 16 for model summary. Participants were more likely to choose the knowledge option in response to the Gettier condition than in response to the ignorance condition. The likelihood of choosing knowledge was also higher in response to the knowledge condition than in response to the Gettier and ignorance conditions. Thus, participants demonstrated Gettier intuitions as measured by the alternative knowledge probe as well.

Table 16

Alternative Knowledge Probe Model 6 Summary

Parameter or Statistic	Measure
	Knowledge Probe
Fixed Effects	
Intercept	-0.67 (0.38)
Age	0.01 (0.00)**
Gender	-0.09 (0.04)*
Education	-0.01 (0.01)
Compensation	0.22 (0.04)***
Condition: Ignorance	-1.18 (0.05)***
Condition: Knowledge	0.41 (0.04)***
Random Effects	
Site	< 0.001
Participant	< 0.001
Vignette	0.628
AIC	16332.68

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest an increased likelihood of choosing knowledge. For the condition variable, Gettier was the comparison group. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.
* $p < .05$; ** $p < .01$; *** $p < .001$

Measurement Characteristics

We examined whether condition effects were influenced by measurement characteristics, specifically if the outcome was originally measured on a binary or visual analogue scale. See Table 17 for model summaries. Adding measurement and its interaction with condition to the model predicting knowledge attribution did not produce moderation effects or improve model fit (Model 6D; AIC = 15876.57; Pseudo-R^2 = .21 - .25). Next, we estimated the same model (Model 6D) in predicting judgments of reasonableness (AIC = 7041.29). While this model (Pseudo-R^2 = .02 - .07) performed better than Model 6, the interactions between condition and measurement type were not significant. Finally, we estimated a model that included an interaction between condition and measurement type predicting luck attributions (Model 6D, AIC = 15862.09). This model (Pseudo-R^2 = .14 - .16) performed better than Model 6 and revealed an interaction effect for the Ignorance condition in comparison to the Gettier condition.

Table 17

Measurement Moderation Model (6D) Summaries

Parameter or Statistic	Measure		
	Knowledge	**Reasonableness**	**Luck/(In)ability**
Fixed Effects			
Intercept	-0.23 (0.41)	1.57 (0.28)***	-0.15 (0.38)
Age	0.00 (0.00)	-0.01 (0.00)	0.00 (0.00)*
Gender	-0.08 (0.04)*	-0.18 (0.07)*	0.01 (0.04)
Education	-0.02 (0.01)*	0.06 (0.01)***	-0.03 (0.01)***
Compensation	0.02 (0.04)	0.25 (0.07)***	-0.18 (0.04)***
Condition: Ignorance	-1.27 (0.14)***	-0.34 (0.18)	1.42 (0.13)***
Condition: Knowledge	0.59 (0.12)***	0.48 (0.22)*	0.82 (0.12)***
Measurement: VAS	0.06 (0.09)	0.35 (0.15)*	-0.13 (0.10)
Measurement X Ignorance	-0.06 (0.15)	-0.07 (0.20)	-0.45 (0.14)***
Measurement X Knowledge	0.02 (0.13)	-0.06 (0.24)	0.14 (0.13)
Random Effects			
Site	< 0.001	0.086	0.072
Participant	< 0.001	0.070	0.010
Vignette	0.669	0.277	0.613
AIC	15876.57	7041.30	15862.10

Note. Estimates and their standard errors, in parentheses where applicable, are provided for each variable in the model. Positive values suggest increased likelihood of knowledge attributions, reasonableness judgments, or attributions to (in)ability. For the condition variable, Gettier was the comparison group. For the measurement variable, binary was the comparison group. Full model statistics can be found at https://osf.io/8ut6e/ in the analysis folder.
* *p* < .05; ** *p* < .01; *** *p* < .001

Condition affected the likelihood of luck attributions on responses to the binary measure, $\chi^2(2) = 120.98$, $p < .001$. Participants were more likely to attribute outcomes to luck in the Gettier condition ($\hat{p}$ = .68) than in the ignorance condition ($\hat{p}$ = .37; Cramér's V = .31, 95% CI

[.26, .37], $\chi^2(1) = 118.14$, $p < .001$). Participants were also more likely to attribute outcomes to luck in the Gettier condition ($\hat{p} = .55$) than in the knowledge condition ($\hat{p} = .37$; Cramér's $V = .18$, 95% CI [.13, .24], $\chi^2(1) = 41.00$, $p < .001$).

Condition similarly affected luck attributions as measured by the VAS, $\chi^2(2) = 454.78$, $p < .001$. Participants were more likely to attribute outcomes to luck in the Gettier condition ($\hat{p} = .66$) than in the ignorance condition ($\hat{p} = .45$; Cramér's $V = .22$, 95% CI [.20, .24], $\chi^2(1) = 341.27$, $p < .001$). Participants were also more likely to attribute outcomes to luck in the Gettier condition ($\hat{p} = .66$) than in the knowledge condition ($\hat{p} = .45$; Cramér's $V = .22$, 95% CI [.20, .24], $\chi^2(1) = 345.90$, $p < .001$). The effect size of the difference between the Gettier and knowledge conditions was smaller when attributions to luck were measured continuously, but the confidence intervals of the continuous measure effect sizes overlapped with those produced by the binary measure.

Gettier Scores

Finally, at the request of a reviewer, we compared the rates of knowledge attribution across the Gettier and knowledge conditions by examining so-called Gettier scores. Starmans and Friedman (2020) devised this approach to account for baseline skepticism in comparing differences in knowledge attribution according to condition across subsamples. Gettier scores are calculated by dividing the percentage of participants who attribute knowledge in the Gettier condition by the percentage of those who attribute knowledge in the knowledge condition. Using the values from Model 6 (see Table 7), we computed a Gettier score of 75.47, which suggests that participants on average attributed knowledge in response to the Gettier condition 75.47% as often as they did in response to the knowledge condition. Considering just the Darryl vignette data for participants who responded to it first (i.e., the "direct" replication), yielded a Gettier score of 80.98. These scores highlight the somewhat similar rates of knowledge attribution across the two conditions.

Discussion

Past cross-cultural research has suggested that non-philosophers may rely on a shared epistemic intuition (i.e., a core folk epistemology) that leads them to deny knowledge to protagonists in Gettier-type cases more often than to protagonists in cases of justified true belief, thereby demonstrating Gettier intuitions (e.g., Machery et al., 2017b). In the present research, we examined the prevalence of Gettier intuitions in counterfeit-object Gettier-type cases by replicating and extending Experiment 1 of Turri et al. (2015). Our international multisite study employed three counterfeit-object Gettier vignettes to compare how participants attribute knowledge to protagonists in Gettier, knowledge, and ignorance vignette conditions. Overall, participants demonstrated Gettier intuitions. That is, they were more likely to attribute knowledge to protagonists in standard cases of justified true belief (i.e., the knowledge conditions) than in special cases of justified true belief in which protagonists formed a true belief based on a true observation of an authentic object despite the presence of a salient but failed threat to their ability to detect its authenticity (i.e., the Gettier conditions). This result did not correspond to that found by Turri et al., who failed to detect a difference in knowledge attribution between these two conditions. Notably, the size of the Gettier intuition effect varied by vignette in our research; the Darrel vignette from the original study produced the smallest effect size, one similar to that we calculated using the non-significant result from the original study, and the Emma vignette produced the largest. However, few participants attributed knowledge to Emma regardless of epistemological condition. Our results align with research that suggests that participant perceptions of the protagonist contribute to differences in knowledge attribution rates in Gettier intuition research (e.g., Disher et al., 2021).

Knowledge Attribution

Our results did not correspond to those found by Turri et al. (2015) in a substantive way. In the original study, participants who read the Gettier version of the "Darrel/Squirrel" vignette attributed knowledge to the protagonist at higher rates than those who read the ignorance version. However, the rates of knowledge attribution did not differ between participants in the Gettier and knowledge control conditions. Although we similarly found a large difference between the Gettier and the ignorance conditions in our replication, our analyses also revealed a difference in rates of knowledge attribution between the Gettier and knowledge conditions (i.e., the Gettier intuition effect). This discrepancy could be explained by the low power of the original study (i.e., *N* = 135 in a between-participants design with three levels). Indeed, the original authors suspected that their experiment may have failed to demonstrate a difference between these two conditions due to insufficient power (personal communication with Turri, March 10, 2018). To further examine this possibility, we estimated an effect size for their original analysis for comparison purposes. While non-significant, the original effect (OR = 2.00, 95% CI [0.77, 5.21]) was similar in magnitude to the one we found in our analyses (OR = 1.86, 95% CI[1.78, 1.94). Thus, while we did not replicate Turri et al.'s (2015) null result, they potentially could have also found a significant effect with a sufficiently powered experiment.

Despite this similarity in effect sizes, we argue that our findings do contradict Turri et al.'s (2015) conclusion that "a salient but failed threat to the truth of a judgment does not significantly affect whether it is viewed as knowledge" (p. 381). Given our evidence that participants demonstrated Gettier intuitions for two other similar counterfeit-object Gettier-type cases, which also featured failed threats to the truth of a judgment, we challenge their claim that knowledge attributions are insensitive to such threats. In this way, our results best align with those of other researchers who have found similar effects and concluded that protagonists with luckily true beliefs are less likely to elicit knowledge attributions than protagonists in clear cases of knowledge (Colaço et al., 2014; Machery et al., 2017b; Nagel, San Juan, et al., 2013).

Changes in the methods, design, and analytic approach may also account for differences between our results and those of Turri et al. (2015). One major difference between our replication and the original study was the inclusion of two additional vignettes as part of a within-participants design. The inclusion of these unique stimuli and design features changed the context of the task and may explain some results discrepancies. Unlike in the original study that had a between-participants design, participants in our study responded to all three conditions randomly matched to each vignette in a single experimental session; therefore, participants' responses to a vignette condition may have anchored or led to contrast effects on responses to subsequent vignette conditions. However, participants in the present research were more likely to attribute knowledge to protagonists in the knowledge control condition than in the Gettier condition across all three vignettes, including the one used by Turri et al. (2015). In fact, our exploratory analysis of the Darrel vignette that closely approximated Turri et al.'s original analysis found evidence for Gettier intuitions among participants who responded to that vignette first. Further, and likely because participants were presented with the vignette-condition combinations in a random order, contextual order effects were minimal, and order did not interact with condition in predicting outcomes (see https://osf.io/uz8te).

Prior research on epistemic intuitions has demonstrated the presence of Gettier intuitions among non-philosophers (e.g., Colaço et al., 2014; Nagel, San Juan, et al., 2013) and across cultures and geographic regions (e.g., Machery et al., 2017a, 2017b). Specifically, the limited research using counterfeit-object Gettier-type cases has found that people are generally less likely to attribute knowledge to a protagonist when their true and justified belief is formed on the basis of misleading evidence than in a parallel case when the true and justified belief is formed on the basis of clear evidence (e.g., Weinberg et al., 2001; Nagel, San Juan, et al., 2013).

In the present research, participants likewise demonstrated Gettier intuitions in these cases across different geographic regions. These intuitions were detected on a variety of measures, and knowledge attributions were only minimally (but not meaningfully) affected by participant characteristics such as gender, age, and years of education. Although prior research has suggested that differences in knowledge attribution may depend on how participants are asked whether a target has knowledge (e.g., Machery et al., 2017b; Nagel, San Juan et al., 2013), we found the same pattern of results on the continuous measure, the original binary measure (*knows* vs. *only believes*), and the alternative knowledge attribution measure (*knows* vs. *feels like they know but does not know*). Thus, the present research supports the view that non-philosophers generally demonstrate Gettier intuitions and may to some extent rely on a shared core folk epistemology (i.e., intuitions about knowledge) when assessing the knowledge of others. However, our findings using counterfeit-object Gettier-type cases may not generalize broadly to other categories of Gettier-type cases (e.g., reliabilist, apparent evidence), which may elicit different epistemic intuitions. Further, a notable number of participants (43.41%; see Table 7) denied knowledge to protagonists even in clear cases of justified true belief; thus, this supposed "core folk epistemology" is not universally shared. After accounting for such baseline skepticism, participants on average attributed knowledge in response to the Gettier condition

75.47% as often as they did in response to the knowledge condition. While Gettier protagonists were deemed ignorant more often than not, Gettier intuitions were by no means common.

Ancillary Findings

Reasonableness judgments

According to the Justified True Belief (JTB) account of knowledge, protagonists must be perceived as having met all three criteria (i.e., justification, truth, and belief) to be attributed knowledge (Jacquette, 1996). To test whether Gettier-type challenges to standard justified true beliefs produce different rates of knowledge attribution in counterfeit-object cases, we evaluated whether conditions were perceived as having met the appropriate criteria for the Justified True Belief analysis of knowledge. In the present research, the vignette comprehension questions served as the belief criteria by ensuring that participants could report that protagonists held the relevant belief. The truth of the protagonists' belief varied by condition (i.e., only the protagonist in the ignorance condition held a false belief). The reasonableness judgment measure assessed whether participants judged the protagonists' beliefs to be justified (i.e., reasonable). In the original study, Turri et al. (2015) found no difference between the three epistemological conditions in participants' reasonableness judgments (i.e., how reasonable the participant rated the protagonist for holding their given belief). The authors interpreted this null result as evidence that differences in knowledge attribution could not be explained by differences in judgments of the protagonists' reasonableness by condition.

In the present research, condition did minimally affect whether participants judged protagonists as reasonable. Participants were more likely to judge protagonists in the Gettier conditions as reasonable than protagonists in the ignorance conditions. They were also more likely to judge protagonists as reasonable in the knowledge condition than in the Gettier condition. While we did detect small differences in judgments of reasonableness between conditions, the vast majority of participants responded that protagonists were reasonable in all conditions. Thus, participants generally perceived the protagonists as being justified in their belief regardless of vignette or condition. Further, the high statistical power of our study allowed us to detect very small effects of condition on reasonableness judgments. Such small differences were unlikely to have had much impact on knowledge attributions; however, we did not directly examine this causal pathway.

Luck Attributions

Prior research suggests that attributions of true beliefs to luck may moderate the extent to which Gettier intuitions are demonstrated; when Gettier protagonists are perceived as lucky (as opposed to able), the likelihood they are denied knowledge appears to increase (Turri, 2016, 2017). In the present research, participants attributed outcomes to luck more frequently in the Gettier condition than in the other two conditions. As expected, we found a negative relationship between the likelihood of luck attributions and the likelihood of knowledge attributions. However, we failed to find evidence that the magnitude of the Gettier intuition effect was moderated by luck attributions. While results suggested a moderating effect of attributions to luck or (in)ability on the difference between the ignorance and Gettier conditions, the difference in knowledge attributions between the Gettier and knowledge conditions did not differ according to whether participants attributed truth outcomes to luck or to ability. Seemingly, the likelihood of Gettier intuitions did not depend on participants attributing the protagonist's true belief to luck. However, the stark differences in luck attributions between vignettes may have dampened moderation effects that could have been found if we had examined a single scenario.

Differences between Vignettes

In prior research, Gettier intuitions have been investigated using a variety of different Gettier-type cases and methodologies. Across the types of Gettier-type cases (e.g., "replacement-by-backup", "counterfeit-object", "authentic-evidence", "apparent-evidence"), research results often contradict one another. Previous research suggests that heterogeneous findings can sometimes be explained by methodological features of the research, such as the stimuli used (e.g., Kenny & Judd, 2019; Landy et al., 2020). In line with this view, the present research found that vignette moderated the effect of condition on all considered dependent measures to varying degrees.

Despite possessing the same epistemological structure, the three tested vignettes produced different rates of knowledge attribution, both overall and according to condition (see Figure 3). Participants attributed considerably less knowledge to Emma in the fake diamond vignette than to the protagonists in the other two vignettes. These findings align with prior research that provided evidence for the prevalence of Gettier intuitions using the "Emma/Diamond" vignette (Disher et al., 2021; Nagel, San Juan, et al., 2013; Powell et al., 2015). For example, while Powell et al. (2015) found different rates of knowledge attribution among participants in the ignorance, Gettier, and knowledge conditions, few participants attributed knowledge to Emma overall (e.g., just 25% of those in the knowledge condition). However, in Experiment 4 reported by Turri et al. (2015), participants in the knowledge condition of a similar "Emma/Diamond" vignette attributed knowledge at a similar rate (90%) to those in the Gettier condition involving a failed threat (83%). Notably, the epistemological structure of the Gettier condition in Turri et al. (2015, Experiment 4) differed from that employed in the present research. Thus, the strength of Gettier intuitions we observed for the Emma vignette appears to cohere with prior research.

Knowledge attributions for the Gerald vignette were overall more split compared to the other two vignettes. However, making comparisons to past empirical research that used the "Gerald/House" vignette is difficult given that prior studies that have used it relied on very different methodology and study materials (Colaço et al., 2014; Disher et al., 2021; Swain et al., 2008; Ziółkowski, 2016). Some researchers have found differences in knowledge attributions between Gettier conditions and knowledge conditions for this vignette, albeit using different methodologies (Colaço et al., 2014; Disher et al., 2021; Ziółkowski, 2016). Thus, in line with our findings, most research using the Gerald vignette has found evidence for Gettier intuitions. Besides the original study (Turri et al., 2015) and the present replication, only one other study (Disher et al., 2021) has employed the "Darrel/Squirrel" vignette to our knowledge; Disher et al. did not find evidence for Gettier intuitions in response to this vignette, although they used a different name.

One reason why our vignettes elicited different rates of knowledge attribution may relate to perceptions of luck; vignette moderated the effect of condition on both knowledge and luck attributions. For luck attributions, differences between the Gettier and ignorance conditions were considerably smaller for the Darrel and Gerald vignettes (Cramér's $V = .02$ and Cramér's $V = .20$, respectively) than for the Emma vignette (Cramér's $V = .50$), and the luck attribution differences between the Gettier and knowledge conditions were also smaller for the Darrel and Gerald vignettes (Cramér's $V = .17$ and Cramér's $V = .20$, respectively) than for the Emma vignette (Cramér's $V = .32$). Overall, Emma's outcomes were attributed most often to luck and Darrel's outcomes were attributed most often to ability. Thus, the reason why vignettes differed in their overall level of both knowledge and luck attributions may relate to the perceived characteristics of the target protagonist or their situation. In further support of this view, a separate extension of the present research manipulated the gender of the target protagonist and found that a female protagonist’s knowledge outcome was more likely to be attributed to luck (as opposed to ability) than that of a male protagonist’s across all conditions and vignettes (Disher et al., 2021). Thus, the gender of the protagonist may have potentially served as a cue that participants used to assess the ability of a protagonist when deciding whether or not they possessed knowledge.

However, in the present research, differences in the results produced by the Emma vignette in comparison to the other vignettes cannot easily be attributed to protagonist gender alone. Other factors unique to the Emma vignette may also partially explain the differences in response rates across vignettes. For instance, the Emma vignette introduced skeptical pressure in ways the other two vignettes did not. Specifically, participants in all conditions read that Emma, “could not tell the difference between a real diamond and a cubic zirconium fake,” suggesting a lack of expertise and subsequent knowledge. In an extension carried out by collaborators (Larkin & Andreychik, 2019; see also Appendix E), an additional vignette that manipulated the perceived expertise level of protagonists (i.e., expert or novice) and the condition (i.e., knowledge, Gettier, or ignorance) was tested as part of our data collection in a fully between-participants design. Their results demonstrated that the perceived expertise of protagonists affected knowledge attribution rates; protagonists with high expertise were more likely to be attributed knowledge than those with low expertise. Given that the Darrel vignette features a protagonist that is described as being an ecologist (i.e., an expert) and the Emma vignette features a protagonist that is described as not able to evaluate whether a diamond is authentic (i.e., not an expert), differences in attribution rates between these vignettes may be due to their perceived level of expertise.

Finally, the Emma vignette also featured a scenario with which most participants were more likely to have personal experience (i.e., shopping). In contrast, the Darrel vignette featured a scenario with which most participants were less likely to have personal experience (i.e., ecological research). Nagel, San Juan, et al. (2013) argued that epistemic egocentrism, or the tendency of people to evaluate others as though they know what we know (Birch, 2005; Birch & Bloom, 2007; Camerer et al., 1989; Nickerson, 1999), may play a substantial role in how participants evaluate the knowledge of others. If participants have differing levels of pre-existing knowledge about vignette scenarios, they may be differently equipped to evaluate protagonists in each scenario based on their assumed shared knowledge. Perhaps, participant familiarity with the context of shopping in the Emma vignette allowed participants to consider ways in which she could have better evaluated her belief. Participant familiarity with the context of ecological research was likely comparatively low; they may not have generated alternative approaches for Darrel to evaluate his belief. Because we did not manipulate these features of the tested vignettes, such interpretations remain speculative. Parsing out the effects of these different sources of stimulus variation would be a valuable aim for future research.

Implications

Previous research on epistemic intuitions has primarily focused on whether lay people deny knowledge to targets in philosophical problems based on the epistemological structure of the problem. Secondarily, research has investigated whether lay denials of knowledge in these sorts of problems differ based on the identity of the rater/participant (e.g., participants' gender, class, or culture). Our results demonstrate that epistemological structure and participant identity alone cannot fully account for the rate at which people deny or attribute knowledge. Even standard cases of justified true belief were attributed knowledge at different rates between these vignettes.

In the present research, all scenarios represented the same type of Gettier-type case (i.e., counterfeit-object cases) and thus featured the same epistemological structure. If people's epistemic intuitions rely only on all of the same epistemological criteria (e.g., justification, truth, and belief), then they should have denied knowledge similarly across these scenarios as a function of whether those criteria were met. Instead, our results suggest that people attribute knowledge in ways that deviate from these theoretical expectations. Specifically, characteristics of the protagonists and situations presented in the vignettes seem to moderate attributions of knowledge.

While participants' knowledge attributions may have been sensitive to the nuances of the tested vignettes, the way in which participants attributed knowledge was fairly straightforward. Most participants attributed knowledge on a continuous visual analogue scale which allowed for, but did not reveal, considerable variability in the degree of knowledge attributed to the protagonist. Instead, participants responded in a clearly binary manner as revealed by the bimodal distribution of the knowledge variable: Protagonists were generally perceived as either having knowledge or not. These findings in and of themselves demonstrate that people make judgments about knowledge in a very dichotomous manner.

Pedagogical Considerations

As a partnership between the PSA and CREP, this project had a central goal of serving a pedagogical function with support through the PSA's network and resources. Experiment 1 from Turri et al. (2015) was selected by the CREP team as a study that was feasible for students to directly replicate; the original study had relatively simple materials (i.e., three variations of one "Darrel" vignette), measurements (i.e., dichotomous "knows/believes" judgments), and analyses (i.e., chi-square goodness-of-fit tests). In the process of submitting and revising a Registered Report for the study, the materials, measurements, and analyses all became more complex and, importantly, more useful for the underlying empirical questions than the original. However, we observed some trade-offs between rigor and pedagogy because of this increase in complexity.

In a typical CREP project, students prepare their materials and OSF pages, submit their pages for initial review, collect data, clean and analyze data, interpret their results, and submit their pages for final review. The increase in design complexity resulted in the need for centralized data collection to guarantee adherence to the randomization and counterbalancing procedures. Instead, students worked with the project administration team to incorporate their own information (e.g., informed consent, compensation) into the centralized survey where needed. The increased analytic complexity meant that students (and instructors) faced challenges in completing their site-level analyses. The majority of undergraduate and masters level students have likely not been trained in mixed ANOVA or multilevel modeling.

For this project, students generally did not prepare their own materials or analyze their own site's data (see the "Site Level Analysis" column in Table 1). However, most or all students completed many traditional CREP steps: creating OSF accounts and following instructors to create study pages for their site, recording videos of the study procedures, posting all materials including ethics approval, requesting reviews, and revising projects as necessary. All teams with data included in the present study completed at least these minimum requirements; some teams did more than the minimum required, including the evaluation of extension hypotheses. In large part, however, teams just completed the minimum requirements.

In general, we believe that student contributors may have received less training by participating in this project than they would have during a typical CREP project. We have planned a follow-up survey to assess self-reported learning among student collaborators. While we can compare the results of that survey to similar surveys following other CREP projects, we cannot determine whether participation in the project would have produced different learning outcomes for students had it been implemented as originally planned.

The tradeoffs between the scientific and pedagogical aims of this study had other consequences. Our attempt to provide flexibility for teams resulted in data loss and energy-, time-, and resource-draining data processing procedures. For instance, some contributors requested the ability to prepare their own project materials via Qualtrics and, in consultation with the Registered Replication Report editor, we decided to support the pedagogical goals of those researchers. This effort to allow for experiential learning while adhering to the approved methods and analysis plan led to complications. Data from some of the teams who administered a Qualtrics survey proved unusable due to lack of adequate documentation.

If this had been a purely PSA study, then students would presumably have had fewer opportunities to participate in educational activities like using the Open Science Framework or communicating with reviewers before data collection. Students also would have had less flexibility in data collection methods and extension variables. On the other hand, data processing and documentation would have been much easier. If we were only interested in addressing the empirical questions of this research, or if we were only interested in training students how to do replications or research, our approach would not have been appropriate. We exchanged time, resources, and energy for the opportunity to satisfy both empirical and pedagogical goals. Creative strategies, such as requiring students to prepare materials on their own prior to being given access to the centralized data collection link, may satisfy the needs of both pedagogy and rigor in future large-scale collaborations.

Despite these tradeoffs, we would recommend doing big team science with student researchers in the future. Likely, some of our challenges may have been less pronounced without a Registered Report process that placed a priority on the empirical question and resulted in a complicated design. At the very least, the students who collaborated as researchers on this project learned about preregistration, Registered Reports, and the Open Science Framework. General research literacy can be improved by learning about these practices and, for those students who will continue to do research in graduate school or as part of their profession, incorporating these practices into their toolkit at an early stage may improve the rigor and transparency of their future contributions (Pownall et al., 2023).

Limitations

Though the present research represents the largest multisite empirical study of Gettier intuitions to date and was conducted across multiple geographic regions using multiple minimally matched stimuli, our conclusions are limited by (1) inconsistencies in data documentation and collection, (2) methodological decisions, (3) strict a priori exclusion criteria, and (4) generalizability.

Given the pedagogical goals of this project, trade-offs between research quality and accessibility to students were made at various stages of the project that led to inconsistencies in data documentation and collection. Exceptions to the accepted protocol were granted for several student teams (e.g., some teams implemented the study independently in Qualtrics rather than using the vetted SoSciSurvey survey). Thus, some of the samples collected as part of this project were excluded due to data quality concerns. However, despite losses in data due to these exclusions, permitting flexibility in data collection allowed for more students to experience being part of a large multisite research project that enriched their research education.

Methodological complications further limit our results. The original experiment used binary response options for the dependent measures; as planned, we implemented visual analogue scales instead. This difference may have impacted the results that we found before and after converting those continuous responses to a binary format. Exploratory analyses suggested that a binary knowledge measure, a randomly assigned alternative implemented by some teams, did not produce meaningfully different results from those we obtained using the dichotomized continuous knowledge measure. Further, using the untransformed continuous measures in analyses produced a similar pattern of results as those we reported (see https://osf.io/nvfbm/). Still, our findings may have been different if all participants were asked to respond to the knowledge question in a response format that better reflects the binary way in which people appear to make these kinds of determinations. Additionally, the exploratory luck vs. ability measure was originally planned to be a single question that required two responses. We changed how the question was displayed to alleviate participant confusion, but this deviation may have affected responding. Finally, we were unable to use two of the planned test setting covariates (i.e., online vs. in person and individually vs. in a group) in our analyses due to unforeseen challenges in data collection (e.g., changing modalities due to the COVID-19 pandemic). The inclusion of these variables may have impacted our results.

The large number of participant exclusions is another potential limitation of this research. According to our strict a priori exclusion criteria, many participants were excluded because they responded incorrectly to at least one of the vignette comprehension questions (46.36% of participants met this criteria), had missing or invalid data for age (22.44% of participants met this criteria), and/or did not respond the language proficiency question or reported low proficiency (22.17% of participants met this criteria). These three exclusion criteria resulted in nearly half of participants being excluded from analyses. Failed vignette comprehension checks accounted for most of the exclusions, likely due to inattention or the intellectually challenging content. However, the direct replication analysis using data from only the Turri/Squirrel vignette that only excluded participants who got the corresponding comprehension question wrong closely mirrored our primary findings. Additional exploratory analyses excluding participants who failed a specific comprehension question, rather than employing listwise exclusions, demonstrated a similar pattern of results (see https://osf.io/nvfbm/). Further, although nearly half the participants were excluded, potentially limiting the generalizability of our results, our strict criteria arguably increased the validity of our findings by including only those who understood the scenarios.

Comprehension exclusion rates have varied widely in previous Gettier intuition investigations (e.g., 2% - 47%; Machery et al., 2017a; Starmans & Friedman, 2012), but those studies used between-participants designs in which participants responded to a single scenario. Our relatively high rates of comprehension exclusions (i.e., 46%) may have resulted from our listwise exclusion of participants if they responded incorrectly to any one of the three vignettes' comprehension questions. However, other cross-cultural studies in this domain have produced similar comprehension exclusion rates with between-participants designs (e.g., 47%, Machery et al., 2017a). Perhaps cultural variation in conceptual familiarity or linguistic forms reduced comprehension or memory of the tested vignettes (see Machery et al., 2017a). Regardless, according to a review of Gettier intuition studies (Popiel, 2016), participant exclusions typically have no effect on study results. Still, we cannot easily draw conclusions about lay people's epistemic intuitions given their difficulty engaging with our scenarios. This potential limitation may broadly apply to the field of experimental philosophy. Often, experimental philosophy research introduces participants to highly abstract and intricate scenarios with underlying assumptions that lay people do not accept or struggle to understand (e.g., Bergenholtz et al., 2021; Murray et al., 2022).

We chose to not execute an additional planned exclusion, which would have removed participants from sites where teams did not receive a CREP completion certificate. As discussed in the Method section, we decided not to exclude data from teams that were approved for data collection and used the centralized survey even if they did not receive certificates. Requiring completion of the remaining pedagogical tasks would have further reduced our sample size without meaningfully increasing quality assurance. Further, as previously explained, implementing this additional exclusion criteria did not substantively impact results (see https://osf.io/nvfbm).

Lastly, because most of our participants were drawn from university samples, our findings may not generalize beyond the small subset of educated, socioeconomically advantaged young adults—at least those able to pass comprehension checks (for evidence regarding socioeconomic differences, see Nichols et al., 2003; for educational differences, see Starmans & Friedman, 2012; for age differences, see Colaço et al., 2014). However, our results indicated that age and years of education had only very small associations with knowledge attribution rates that were not robust to changing model specifications. Also, given that our sample of Gettier-type cases from the epistemology literature was limited to counterfeit-object scenarios, inferences made from our findings should be applied only to intuitions in that subset of Gettier-type cases. Other forms of Gettier-type cases (e.g., evidence replacement cases) may produce different epistemic intuitions. For example, prior literature has demonstrated that participants are less likely to attribute knowledge to protagonists in Gettier-type cases that present “apparent” evidence (e.g., Turri, 2013) and more likely to attribute knowledge in cases that present “authentic” evidence (e.g., Starmans & Friedman, 2013). We have no reason to believe that the results presented in this paper were dependent on other characteristics of the participants, materials, or context (Simons et al., 2017).

Conclusion

Turri et al. (2015) interpreted their original findings as supporting the view that a salient but failed threat to the truth of a judgment does not affect whether it is viewed as knowledge. The results from this RRR suggest that this view should be amended. Contrary to Turri et al.'s claim, our participants attributed knowledge significantly more often to protagonists in standard justified true belief cases than in counterfeit-object Gettier-type cases. However, we did observe a smaller Gettier intuition effect in the vignette used in the original study than in the other vignettes we employed. Overall, our results suggest that attributions of knowledge may be affected by contextual characteristics unrelated to the knowledge criteria met by protagonists, such as perceptions about protagonists’ ability and expertise. Future research on epistemic intuitions should focus on identifying the moderating role of contextual characteristics to better understand the conditions necessary for people to attribute knowledge to others.

Author Contributions

B. Hall, J. Wagge, C. R. Chartier, M. J. Brandt, J. E. Grahe, H. IJzerman, C. Schild contributed to the conceptualization of the project.

B. Hall, K. Schmidt, J. Wagge, H. Moshontz wrote the original draft of this report.

B. Hall, K. Schmidt, J. Wagge, S. C. Lewis, S. Weissgerber, F. Kiunke, G. Pfuhl, S. M. Stieger, U. S. Tran, K. Barzykowski, N. Bogatyreva, M. Kowal, K. Massar, F. Pernerstofer, P. Sorokowski, M. Voracek, C. R. Chartier, M. J. Brandt, J. E. Grahe, A. A. Özdoğru, M. R. Andreychik, S. Chen, T. R. Evans, C. Hautekiet, H. IJzerman, P. Kačmár, A. J. Krafnick, E. D. Musser, E. Vergauwe, K. M. Werner, B. Aczel, P. Arriaga, C. Batres, J. L Beaudry, F. Cova, L. D. Cramblet Alvarez, G. Feldman, H. Godbersen, J. Gottfried, G. J. Haeffel, A. Hartanto, C. Isloi, J. P. McFall, M. Milyavskaya, D. Moreau, E. Nosáľová, K. Papaioannou, S. Ruiz-Fernandez, J. Schrötter, D. Storage, K. Vezirian, L. Volz, Y. J. Weisberg, Q. Xiao, S. Ďurbisová, D. Awlia, H. W. Branit, M. R Dunn, A. Groyecka-Bernard, R. Haneda, G. Kalistová, J. Kielinska, C. Kolle, P. Lubomski, A. M. Miller, M. J. Mækelæ, M. Pantazi, R. R Ribeiro, R. M. Ross, A. Sorokowska, C. L. Aberson, X. Alexi Vassiliou, B. J. Baker, M. Bognar, C. Cong, A. F. Danvers, W. E. Davis, V. Dranseika, A. Dumbravă, H. Farmer, A. P. Field, P. S. Forscher, A. Graton, N. Hajdu, P. A. Howlett, R. Kabut, E. M. Larsen, S. T. H. Lee, N. Legate, C. A. Levitan, N. Levy, J. G. Lu, M. Misiak, R. Elana Morariu, J. Novak, E. Pronizius, I. Prusova, A. S. Rathnayake, M. O. Romanova, J. P. Röer, W. M. Sampaio, C. Schild, M. Schulte-Mecklenbeck, I. D Stephen, P. Szecsi, E. Takacs, J. N. Teeter, E. Hugh Thiele-Evans, J. Valeiro-Paterlini, I. Vilares, L. Villafana, K. Wang, R. Wu, S. Álvarez-Solas, H. Moshontz, E. M. Buchanan edited, reviewed, and revised the draft of this report.

B. Hall, J. Wagge, S. C. Lewis, S. Weissgerber, F. Kiunke, C. Hautekiet, A. J. Krafnick, H. Moshontz, E. M. Buchanan contributed to the data curation.

B. Hall, K. Schmidt, F. Kiunke, G. Pfuhl, U. S. Tran, M. Voracek, M. J. Brandt, P. S. Forscher, H. Moshontz, E. M. Buchanan contributed to the analysis of the data.

B. Hall, K. Schmidt, J. Wagge, S. C. Lewis, S. Weissgerber, G. Pfuhl, S. M. Stieger, K. Barzykowski, K. Massar, P. Sorokowski, C. R. Chartier, M. J. Brandt, J. E. Grahe, A. A. Özdoğru, M. R. Andreychik, S. Chen, T. R. Evans, H. IJzerman, P. Kačmár, A. J. Krafnick, E. D. Musser, E. Vergauwe, K. M. Werner, B. Aczel, P. Arriaga, C. Batres, J. L Beaudry, L. D. Cramblet Alvarez, G. Feldman, J. Gottfried, G. J. Haeffel, A. Hartanto, C. Isloi, J. P. McFall, M. Milyavskaya, D. Moreau, K. Papaioannou, S. Ruiz-Fernandez, D. Storage, K. Vezirian, L. Volz, Y. J. Weisberg, C. L. Aberson, B. J. Baker, H. Farmer, C. A. Levitan, H. Moshontz, E. M. Buchanan was involved in supervision for this project including supervision of student led teams.

B. Hall, K. Schmidt, J. Wagge, S. C. Lewis, S. Weissgerber, F. Kiunke, G. Pfuhl, S. M. Stieger, U. S. Tran, K. Barzykowski, N. Bogatyreva, M. Kowal, K. Massar, F. Pernerstofer, P. Sorokowski, M. Voracek, C. R. Chartier, J. E. Grahe, A. A. Özdoğru, M. R. Andreychik, S. Chen, T. R. Evans, C. Hautekiet, P. Kačmár, A. J. Krafnick, E. D. Musser, E. Vergauwe, K. M. Werner, B. Aczel, P. Arriaga, C. Batres, J. L Beaudry, L. D. Cramblet Alvarez, G. Feldman, H. Godbersen, J. Gottfried, G. J. Haeffel, A. Hartanto, C. Isloi, J. P. McFall, M. Milyavskaya, D. Moreau, E. Nosáľová, K. Papaioannou, S. Ruiz-Fernandez, J. Schrötter, D. Storage, L. Volz, S. Ďurbisová, D. Awlia, H. W. Branit, M. R Dunn, A. Groyecka-Bernard, R. Haneda, G. Kalistová, C. Kolle, A. M. Miller, R. R Ribeiro, R. M. Ross, A. Sorokowska, W. E. Davis, A. Dumbravă, N. Hajdu, P. A. Howlett, S. T. H. Lee, N. Legate, N. Levy, M. Misiak, J. Novak, I. Prusova, A. S. Rathnayake, M. O. Romanova, I. D Stephen, P. Szecsi, E. Takacs, J. N. Teeter, E. Hugh Thiele-Evans, I. Vilares, L. Villafana, R. Wu, S. Álvarez-Solas, E. M. Buchanan contributed to the data collection and investigation.

B. Hall, J. Wagge, S. Weissgerber, G. Pfuhl, S. M. Stieger, U. S. Tran, C. R. Chartier, M. J. Brandt, J. E. Grahe, A. A. Özdoğru, M. R. Andreychik, S. Chen, H. IJzerman, E. D. Musser, E. Vergauwe, K. M. Werner, F. Cova, V. Dranseika, A. P. Field, H. Moshontz, E. M. Buchanan contributed to methodology for the project.

J. Wagge, S. Weissgerber, G. Pfuhl, N. Bogatyreva, M. Kowal, F. Pernerstofer, C. R. Chartier, A. A. Özdoğru, T. R. Evans, C. Hautekiet, P. Kačmár, F. Cova, H. Godbersen, E. Nosáľová, J. Schrötter, K. Vezirian, Y. J. Weisberg, Q. Xiao, S. Ďurbisová, J. Kielinska, P. Lubomski, M. J. Mækelæ, M. Pantazi, X. Alexi Vassiliou, M. Bognar, C. Cong, A. Graton, R. Kabut, E. M. Larsen, J. G. Lu, R. Elana Morariu, E. Pronizius, J. P. Röer, W. M. Sampaio, M. Schulte-Mecklenbeck, J. Valeiro-Paterlini, K. Wang, H. Moshontz, E. M. Buchanan contributed to resources for this project.

S. Weissgerber, F. Kiunke, A. F. Danvers, E. M. Buchanan contributed survey software and coding for this project.

E. M. Buchanan contributed validation of the analyses.

J. Wagge, G. Pfuhl, Q. Xiao, E. M. Buchanan contributed to the visualization of data and analyses.

Conflicts of Interest

The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

Patrícia Arriaga was supported by UID/PSI/03125/2019 from the Portuguese National Foundation for Science and Technology (FCT).

Pavol Kačmár was supported by project KEGA-079UK-4/2021.

While working on the study and resulting paper, Krystian Barzykowski was supported by the National Science Centre, Poland [No.: 2015/19/D/HS6/00641; 2019/35/B/HS6/00528].

Piotr Sorokowski was funded by IDN Being Human Lab (University of Wrocław).

Erica D. Musser was supported by K23MH117280.

Kaitlyn M. Werner was supported by a Banting Postdoctoral Research Fellowship awarded by the Social Sciences and Humanities Research Council of Canada (SSHRC).

Rafael R. Ribeiro was supported by UID/PSI/03125/2019 from the Portuguese National Foundation for Science and Technology (FCT).

Agnieszka Sorokowska was supported by Scientific Excellence Incubator "Being Human. (University of Wroclaw).

Neil Levy was supported by the Australian Research Council (DP180102384) and the John Templeton Foundation (62631).

Robert M Ross was supported by the Australian Research Council (DP180102384) and the John Templeton Foundation (62631).

Michał Misiak was supported by the scholarship of the National Science Centre (2020/36/T/HS6/00256) and by the Scientific Excellence Incubator of the University of Wroclaw – Being Human.

Supplemental Material

Supplemental materials include the provisionally accepted version of the manuscript (Appendix A), and study materials: vignettes and dependent variable questions (Appendix B); covariate, demographic, and exclusion questions (Appendix C); open response coding instructions and details (Appendix D); and the vignette used in the expertise extension (Appendix E).

Prior Versions

The provisionally accepted Stage 1 version of this manuscript was preregistered (https://osf.io/4bfs7) and can be found in Appendix A. The introduction was restructured and edited for clarity. The methods section was rewritten to increase accuracy and conformity to reporting norms. The analysis plan was revised to correct errors and correspond to features of the data. Additional prior versions have been posted as preprints at https://psyarxiv.com/zeux9/.

References

Bartoń, K. (2020). MuMIn: Multi-Model Inference (1.43.17) [Computer software]. https://CRAN.R-project.org/package=MuMIn

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions.* SAGE Publications. https://doi.org/10.4135/9781412986373

Bergenholtz, C., Busch, J., & Praëm, S. K. (2021). Exclusion Criteria in Experimental Philosophy. *Erkenntnis, 86*(6), 1531-1545. https://doi.org/10.1007/s10670-019-00168-5

Birch, S. A. J. (2005). When Knowledge Is a Curse: Children’s and Adults’ Reasoning About Mental States. *Current Directions in Psychological Science*, *14*(1), 25–29. https://doi.org/10.1111/j.0963-7214.2005.00328.x

Birch, S. A. J., & Bloom, P. (2007). The Curse of Knowledge in Reasoning About False Beliefs. *Psychological Science*, *18*(5), 382–386. https://doi.org/10.1111/j.1467-9280.2007.01909.x

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the likert item responses and other ordinal measures. *International Journal of Exercise Science*, *8*(3), 297–302. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4833473/

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van ’t Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Buckwalter, W., & Stich, S. (2010). Gender and philosophical intuition. SSRN. http://dx.doi.org/10.2139/ssrn.1683066

Camerer, C., Loewenstein, G., & Weber, M. (1989). The Curse of Knowledge in Economic Settings: An Experimental Analysis. *Journal of Political Economy*, *97*(5), 1232–1254. https://doi.org/10.1086/261651

Chartier, C. R., & McCarthy, R. J. (2018). *StudySwap Tutorial* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/wqhbj

Colaço, D., Buckwalter, W., Stich, S., & Machery, E. (2014). Epistemic intuitions in fake-barn thought experiments. *Episteme*, *11*(2), 199–212. https://doi.org/10.1017/epi.2014.7

DePaul, M. R., & Ramsey, W. M. (Eds.). (1998). *Rethinking intuition: The psychology of intuition and its role in philosophical inquiry*. Rowman & Littlefield.

Disher, N. G., Guerra, A. L., & Haeffel, G. J. (2021). Men have ability, women are lucky: A pre-registered experiment examining gender bias in knowledge attribution. *British Journal of Social Psychology*, *60*(3), 808–825. https://doi.org/10.1111/bjso.12443

Dutant, J. (2015). The legend of the justified true belief analysis: The legend of the justified true belief analysis. *Philosophical Perspectives*, *29*(1), 95–145. https://doi.org/10.1111/phpe.12061

Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, *23*(6), 121–123. https://doi.org/10.1093/analys/23.6.121

Grahe, J., Brandt, M., IJzerman, H., & Cohoon, J. (2014). Replication education. *APS Observer*, 27. https://www.psychologicalscience.org/observer/replication-education

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20. https://doi.org/10.1037/h0076157

Jacquette, D. (1996). Is nondefectively justified true belief knowledge?. *Ratio, 9*(2), 115-127. https://doi.org/10.1111/j.1467-9329.1996.tb00100.x

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103,* 54. https://doi.org/10.1037/a0028347

Kenny, D. A., & Judd, C. M. (2019). The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. *Psychological Methods*, *24*(5), 578–589. https://doi.org/10.1037/met0000209

Kim, M., & Yuan, Y. (2015). No cross-cultural differences in the Gettier car case intuition: A replication study of Weinberg et al. 2001. *Episteme, 12,* 355-361. https://doi.org/10.1017/epi.2015.17

Landy, J. F., Jia, M. (Liam), Ding, I. L., Viganola, D., Tierney, W., Dreber, A., Johannesson, M., Pfeiffer, T., Ebersole, C. R., Gronau, Q. F., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. https://doi.org/10.1037/bul0000220

Larkin, K., & Andreychik, M. (2019). Materials. https://osf.io/fh52n/

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods, 7*(1), 19. https://doi.org/10.1037/1082-989x.7.1.19

Machery, E., Stich, S., Rose, D., Alai, M., Angelucci, A., Berniūnas, R., Buchtel, E. E., Chatterjee, A., Cheon, H., Cho, I.-R., Cohnitz, D., Cova, F., Dranseika, V., Lagos, Á. E., Ghadakpour, L., Grinberg, M., Hannikainen, I., Hashimoto, T., Horowitz, A., … Zhu, J. (2017a). The Gettier Intuition from South America to Asia. *Journal of Indian Council of Philosophical Research*, *34*(3), 517–541. https://doi.org/10.1007/s40961-017-0113-y

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa, K., Struchiner, N., Sirker, S., Usui, N., & Hashimoto, T. (2017b). Gettier across cultures. *Noûs*, *51*(3), 645–664. https://doi.org/10.1111/nous.12110

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*(1), 181. https://doi.org/10.1037/0033-2909.113.1.181

Moser, P. K. (2002). *The Oxford handbook of epistemology*. Oxford University Press.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

Murray, S., Dykhuis, E. D., & Nadelhoffer, T. (2022). *Do people understand determinism? The tracking problem for measuring free will beliefs* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/kyza7

Nagel, J., Juan, V. S., & Mar, R. A. (2013). Lay denial of knowledge for justified true beliefs. *Cognition*, *129*(3), 652–661. https://doi.org/10.1016/j.cognition.2013.02.008

Nagel, J., Mar, R., & San Juan, V. (2013). Authentic Gettier cases: A reply to Starmans and Friedman. *Cognition*, *129*(3), 666–669. https://doi.org/10.1016/j.cognition.2013.08.016

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface, 14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Nichols, S., Stich, S., & Weinberg, J. (2003). Metaskepticism: Meditations in ethno-epistemology. *The skeptics*, 227-247. https://doi.org/10.1093/acprof:oso/9780199733477.003.0010

Nickerson, R. S. (1999). How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others. *Psychological Bulletin*, *125*(6), 737–759. https://doi.org/10.1037/0033-2909.125.6.737

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Popiel, M. (2016). *A systematic review of studies using Gettier-Type thought experiments* (Publication No. 1216) [Doctoral dissertation, University of Denver]. Electronic Theses and Dissertations. https://digitalcommons.du.edu/etd/1216

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. https://doi.org/10.1016/j.cognition.2015.03.002

Pownall, M., Azevedo, F., König, L. M., Slack, H. R., Evans, T. R., Flack, Z., Grinschgl, S., Elsherif, M. M., Gilligan-Lee, K. A., Oliveira, C. M., Gjoneska, B., Kanadadze, T., Button, K. S., Ashcroft-Jones, S., Terry, J., Albayrak-Aydemir, N., Dechterenko, F., Alzahawi, S., Baker, B. J., … FORRT. (2023). Teaching open and reproducible scholarship: a critical review of the evidence base for current pedagogical methods and their outcomes. *Royal Society Open Science, 10*(5), 221255. https://doi.org/10.1098/rsos.221255

Satchell, L. (2018). *[In progress] Collaborators for simple add-on "psych of psych" study*. https://osf.io/ywekt/.

Seyedsayamdost, H. (2015). On normativity and epistemic intuitions: Failure of replication. *Episteme*, *12*(1), 95–116. https://doi.org/10.1017/epi.2014.27

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on Generality (COG): A Proposed Addition to All Empirical Papers. *Perspectives on Psychological Science*, *12*(6), 1123–1128. https://doi.org/10.1177/1745691617708630

Simonsohn, U. (2013). Just Post It: The Lesson From Two Cases of Fabricated Data Detected by Statistics Alone. *Psychological Science*, *24*(10), 1875–1888. https://doi.org/10.1177/0956797613480366

Snijders, T. A. B, & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel* (2nd ed.). SAGE Publications. https://doi.org/10.1080/10705511.2013.797841

Sosa, E. (2007). Experimental philosophy and philosophical intuition. *Philosophical Studies*, *132*(1), 99–107. https://doi.org/10.1007/s11098-006-9050-3

Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition*, *124*(3), 272–283. https://doi.org/10.1016/j.cognition.2012.05.017

Starmans, C., & Friedman, O. (2013). Taking 'know' for an answer: A reply to Nagel, San Juan, and Mar. *Cognition*, *129*(3), 662–665. https://doi.org/10.1016/j.cognition.2013.05.009

Starmans, C., & Friedman, O. (2020). Expert or Esoteric? Philosophers Attribute Knowledge Differently Than All Other Academics. *Cognitive Science*, *44*(7), e12850. https://doi.org/10.1111/cogs.12850

Swain, S., Alexander, J., & Weinberg, J. M. (2008). The Instability of Philosophical Intuitions: Running Hot and Cold on Truetemp: THE INSTABILITY OF PHILOSOPHICAL INTUITIONS. *Philosophy and Phenomenological Research*, *76*(1), 138–155. https://doi.org/10.1111/j.1933-1592.2007.00118.x

Turri, J. (2013). A Conspicuous Art: Putting Gettier to the Test. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3643930

Turri, J. (2016). Vision, knowledge, and assertion. *Consciousness and Cognition, 41*, 41–49. https://doi.org/10.1016/j.concog.2016.01.004

Turri, J. (2017). Knowledge attributions in iterated fake barn cases. *Analysis*, *77*(1), 104–115. https://doi.org/10.1093/analys/anx036

Turri, J., Buckwalter, W., & Blouw, P. (2015). Knowledge and luck. *Psychonomic Bulletin & Review*, *22*(2), 378–390. https://doi.org/10.3758/s13423-014-0683-5

Vickstrom, E. R., Shin, H. B., Collazo, S. G., & Bauman, K. J. (2015). *How Well–Still Good? Assessing the Validity of the American Community Survey's English-Ability Question* (Working Paper SEHSD-WP2015-18). U.S. Census Bureau. https://www.census.gov/content/dam/Census/library/working-papers/2015/demo/SEHSD-WP2015-18.pdf

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., IJzerman, H., Legate, N., & Grahe, J. (2019). A Demonstration of the Collaborative Replication and Education Project: Replication Attempts of the Red-Romance Effect. *Collabra: Psychology*, *5*(1), 5. https://doi.org/10.1525/collabra.177

Weatherson, B. (2013). Margins and errors. *Inquiry*, *56*(1), 63–76. https://doi.org/10.1080/0020174X.2013.775015

Weinberg, J. M., Nichols, S., & Stich, S. (2001). Normativity and epistemic intuitions. *Philosophical Topics*, *29*(1), 429–460. https://doi.org/10.5840/philtopics2001291/217

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences, 45*, E1. https://doi.org/10.1017/S0140525X20001685

Ziółkowski, A. (2016). Folk intuitions and the no-luck-thesis. *Episteme, 13*, 343-358. https://doi.org/10.1017/epi.2015.49

Appendix A

Provisionally Accepted Manuscript

Protocol Development

Deviations from Approved Protocol

In this section, we describe all of the deviations that occurred in this study. Below, we also include the full protocol (i.e., methods and analysis plan) as it was approved by AMPPS with the only edits being made for clarity and/or accuracy. When a deviation occurred in our protocol, we include a footnote with information about how and why we deviated from a particular study design plan. Information regarding deviations from our analysis plan is also summarized at the top of that section below.

Changes to Original Manuscript Text

The term “country” was changed to “geopolitical region” in order to accommodate the fact that not all regions included in the data are recognized as countries (i.e., Taiwan). Referring to Taiwan as a country would result in our Chinese collaborators being unable to participate. We originally described comprehension questions as control variables, but this description was inaccurate. Thus, we reworded sentences that described them as controls to accurately reflect that comprehension questions were planned for exclusion purposes. We previously stated that the Gerald vignette was chosen to be included in this study, in part, because it had minimally matched controls available; however, this statement was inaccurate. The controls used in the study from which it was taken were not similar to those used in the Turri et al. (2015) study and ultimately had to be altered to be usable; the primary reason for the inclusion of the Gerald vignette was its prevalence in the epistemics literature.

Lab Recruitment Deviations

In terms of lab recruitment, four main types of deviations occurred: CREP approval process, preregistration of individual collection sites, call for contributors, and the overall time frame for data collection. First, the approval process for onboarding labs was not strictly followed. Some labs were able to collect data without fully finishing the CREP review process. Further, some labs did not preregister their study or did not make their preregistration public on the OSF. We also did not end up asking AMPPS to make additional calls for contributors, given that we were over capacity for onboarding. Lastly, the time frame for collecting data was extended because of complications that arose due to the Covid-19 pandemic.

Approval Process and Pre-registration. Although we originally planned to only include data from labs that had completed all CREP requirements, we decided to include data from student teams that did not receive completion certificates as long as they used the unaltered version of the centralized survey via SoSciSurvey and met all other inclusion criteria. For the purposes of both quality control and educational value, we originally planned to require all participating labs to preregister their own independent direct replication protocol on the Open Science

Framework (OSF). However, some labs did not carry out their pre-registration on the OSF (or did not make them publicly available) because those student teams did not fully complete the project and thus never made their OSF pages public. However, all CREP approved teams were required to create an OSF page (public or private) that was reviewed and approved before being authorized to collect data. Out of the 69 teams that signed up to contribute, 46 of these labs ended up contributing to the final dataset. Of those 46 teams, 22 teams finished the entire CREP review process and thus earned completion certificates.

Call for Contributors and Data Collection Time Frame. We originally stated that AMPPS would make an additional call for contributors. However, due to being over capacity for onboarding new labs, we had concerns about our small administrative team being able to manage more collection sites. Thus, we opted to not make an additional call for contributors. We planned to begin data collection on June 1st, 2019. However, due to the delays in the review process and in initializing the centralized data collection on SoSciSurvey, data collection could not begin until January 2019. Due to complications caused by the COVID-19 pandemic, we extended the deadline to stop data collection from June 1st, 2020, to June 1st, 2021, to accommodate teams struggling to collect data from their typical participant pools. Due to the deadline for stopping data collection being extended through to June 1st, 2021, because of the COVID-19 pandemic, we continued accepting labs until April of 2021 instead of April 2020.

Method and Design Deviations

In terms of Method and Design, deviations occurred in five main areas: Exceptions to restrictions on data collection, site mistakes in informed consent, use of a data template, changes to how variables were measured or used, and changes to how data was included in the overall analyses. First, we allowed exceptions to the rules we set regarding how sites could collect data (i.e., we allowed some sites to use paid survey sites and Qualtrics). Seven sites mistakenly included slightly different language in their informed consent. Sites did not end up using a data template to submit their data to us, in part because the vast majority of sites collected data via SoSciSurvey and thus their data already fit our data template. Next, we made changes to the measurement and use of some variables. Lastly, we said we would only include site data if they met certain criteria; however, we made some exceptions to these rules as long as data integrity could be ensured.

Data Collection Restriction Exceptions. We originally planned to not allow student teams to collect data on paid survey sites, such as MTurk. In order to meet the pedagogical needs of some student teams, one exception was allowed to ensure that students could still participate in the study given that the pandemic made data collection more difficult for some teams (Data collection site AC2060). We ended up allowing one site to use a paid data collection site to gather data for this study due to pedagogical reasons; although, this site was ultimately not included in analyses for other reasons. We originally said that we would not allow sites to collect data using a survey platform other than SoSciSurvey. However, for pedagogical reasons, some sites were allowed to use Qualtrics with their students. Data from these teams were only included in the overall analyses if we had access to their Qualtrics survey, their raw data, and their codebook.

Informed Consent. Eight collection sites mistakenly worded the study purpose in their informed consent materials slightly differently, but none of these mistakes resulted in the experimenters giving away the true purpose of the study and so were not excluded. More details for these deviations can be found on the OSF. The data collection sites that deviated in this way were AC1916, AC1919, AC1918, AC1934, AC1929, AC2063, AC1940, and AC1935.

Data Template. We originally stated that we would require sites to submit their data to us using a standardized data template. However, a data template was ultimately not used due to the vast majority of samples being collected via the centralized data collection via SoSciSurvey.

Changes to Variables. We originally stated that we would measure the luck attribution variable in one question (i.e., [Protagonist name] got the ___ [response 1: right/wrong] answer because of [his/her] ____ [response 2: (in)ability/(good/bad) luck]). However, during survey development on SoSciSurvey, many contributors noted that the way in which we asked this question was confusing and hard to parse. Therefore, to ensure that this question was clear and easy to respond to for participants, we split the question into two parts (i.e., Part 1: [Protagonist name] got the ___ [response 1: right/wrong] answer.". If the participant answered "right", they saw the second part of the question like this: [Protagonist name] got the right answer because of [his/her] ____ [response 2: (ability/good luck]. If the participant answered "wrong", they saw the second part of the question like this: [Protagonist name] got the wrong answer because of [his/her] ____ [response 2: (inability/bad luck].

We originally planned to use two test setting variables as covariates in our analyses: in a group vs individually and online vs in person. However, given that most sites ended up moving to online collection due to the pandemic, most sites collected data online and individually. Thus, we opted to no longer use these variables because they were no longer viable as covariates. While we originally planned to measure whether participants were sampled from a community or from a university participant pool, labs that opted to collect community samples did not add any question to the survey that would allow us to track which participants were from the "community" and which were typical university students. Thus, the only tracking of this we did was asking labs whether they did any community sampling as part of their study.

We said that we would ask participants if the language they were being tested in was their first language. However, due to an oversight, this question was never asked. We originally planned to require each team to translate and code responses to the open-ended exclusion variables. However, due to logistic issues, we opted to instead ask for bilingual volunteers to rate each response (three ratings per response) to determine if the participant met the exclusion criterion (coded as no = 0, maybe = 1, and yes = 2). If the sum of the three coders' ratings of a participant response was four or above, that participant was excluded from the overall analyses.

We originally stated that we would require translators to have experience with test development. However, this requirement turned out not to be feasible, and, thus, we did not enforce it. We also originally said that the exploratory study experience survey would be added to the end of a site's survey if they opted to do so. However, due to a miscommunication that occurred during the creation of the centralized survey on SoSciSurvey, the study experience questionnaire was included at the end of the core survey; thus, all collection sites ended up including this exploratory measure at the end of their survey. Data from this survey were not used in the present study but may be used for a future paper. The ethnicity variable was measured differently by different labs to ensure that the response options properly reflected the geopolitical region to which participants belonged.

Data Inclusion. We originally stated that each site's participant data would be included in the planned multilevel linear regression analyses after a CREP quality check, which would include reviewing raw data files (checked for errors), post-analysis scripts, codebook, cleaned data files (checked for errors), and narrative summary of project findings (compared to data and analysis). However, these materials were not always available for each student team. Thus, the data from each team was only included in the overall analyses if we were able to check raw data files and codebooks for accuracy and errors, which was only an issue for teams that collected data via Qualtrics. Given that we are using multilevel modeling, we only needed to check the quality of raw data files and codebooks for those data not collected via SoSciSurvey to be considered valid and usable. We automatically had access to the raw data and codebook for all samples collected via SoSciSurvey, so we always used those data rather than any data posted to the OSF. Analysis scripts, cleaned data, and narrative summaries were pedagogically important, but had no bearing on the results of the overall analyses reported in this paper.

Lab Recruitment

The Collaborative Replications and Education Project (CREP) has partnered with the Psychological Science Accelerator (PSA: Moshontz et al., 2018) to conduct a large-scale replication that will combine the innovative pedagogical methods of the CREP with the worldwide collaborative open-science network of the PSA. The purpose of the CREP is to address the need for direct replication work in the field of psychology by utilizing the collective power of student research projects. The CREP selects studies (see our OSF page for details; https://osf.io/n5b3w), and teams of students sign up to run replications of these studies. The CREP oversees the quality of these replications to ensure fidelity. Once enough sites have completed replications, the results from teams that have completed projects are collated to get a more accurate estimation of the effect size.

The PSA is an international network of laboratories created to enable and support crowdsourced research projects with a mission to expedite the accumulation of reliable and generalizable evidence in psychological science (Moshontz et al., 2018). The CREP and PSA partnership, therefore, involves the CREP selecting a study, developing materials, and overseeing the quality of the replications using standard CREP procedures while utilizing the existing PSA network to increase participation among labs. Additionally, the PSA has provided support for a variety of components through its extensive network of experts, including lab recruitment, translations, a data release plan, and expertise on logistical differences between geopolitical regions.

After this study was selected for replication, the executive CREP team publicly announced a call for laboratories interested in participating in the study via email and social media (i.e., Twitter, Facebook), with data collection beginning August 1, 2018 (later changed to January 1, 2019). Over the course of one year, 55 labs from 23 geopolitical regions signed up to contribute samples; however, 10 of these labs have since backed out (mostly from USA). As of this submission, 45 labs from 22 geopolitical regions remain committed. For the purposes of both quality control and educational value, we require that all participating labs pre-register their own independent direct replication protocol on the Open Science Framework (OSF).[21] For quality control, these pages must include a video of an experimental session and be approved by our executive team prior to data collection to ensure that each lab meets all standards and procedures set forth in this protocol. Once ethics approvals, protocols, and session video have been approved by the CREP team, contributors may begin data collection starting June 1st, 2019 (depending on date of in principle acceptance).[22] Overall data collection will end on June 1st, 2020[23]. Data release is dependent upon manuscript acceptance, but full data release will be six months after the first ⅔ of the data release.

Protocol Requirements

Sampling plan. Student teams from any geopolitical region are invited to collect samples.[24] Samples may be collected using the subject pools at each team's institution, social media networks, online, or other methods approved by the CREP team and an IRB. Online data collection services that recruit subjects who are then paid for participating, such as Amazon Mechanical Turk (MTurk), were not originally planned be permitted (see below for rationale behind this restriction)[25]. We will also independently collect one large (n = 500) MTurk sample of US participants to include a sample comparable to that in Turri et al. (2015; see details below). Samples may not be drawn from vulnerable populations or any institutions that house them (such as prisons, mental health facilities, etc.). Samples will consist of people over the age of majority in the location of the study, unless a parent or guardian signs a waiver to participate.

Each site will be required to collect data for all three stimulus sets (i.e., "Darrel", "Gerald", and "Emma") with a target sample size of at least 50 participants (although, some sites will attempt

[21] We originally planned that every lab would preregister their study on the OSF; however, some labs did not carry out their pre-registration on the OSF (or did not make them publicly available) because those student teams did not fully complete the project and thus never made their OSF pages public. However, all teams were required to create an OSF page (public or private) that was reviewed and approved before they were authorized to collect data.

[22] Due to the delays in initializing the centralized data collection on SoSciSurvey, data collection could not begin until January 2019.

[23] Due to complications caused by the COVID-19 pandemic, we extended the deadline to stop data collection to June 1st, 2021, to accommodate teams struggling to collect data from their typical participant pools.

[24] The term "country" was changed to geopolitical region in order to accommodate the fact that not all regions included in the data are universally recognized as countries (i.e., Taiwan).

[25] In order to meet the pedagogical needs of some student teams, a few exceptions were allowed to ensure that students could still participate in the study given that the pandemic made data collection more difficult for some teams.

to collect over 100 participants). Our goal is to collect data from at least 50 independent contributors. We are very close to reaching our contributor goal. Each site's participant data will be included in the planned multilevel linear regression analyses after a CREP quality check, which includes reviewing raw data files (checked for errors), post-analysis scripts, codebook, cleaned data files (checked for errors), and narrative summary of project findings (compared to data and analysis)[26].

Although we will sample from many different populations, results from recent multilab studies with mainly student samples (e.g., the ManyLabs studies) suggest that limited heterogeneity may still be an issue (i.e., samples will likely be predominantly white, socioeconomically advantaged, educated, etc.). We attempt to partially address this concern by encouraging contributing sites to collect non-university participants outside of their typical institutions' sampling pool by rewarding sites who do so with higher author order on the post-data Phase 2 manuscript as well as a CREP quality award.

Testing location. Each contributor's test setting will likely differ in one or more ways from the original Turri et al. (2015) study which was completed online using MTurk. To extend the generalizability of this replication, teams may test their samples either in person or online. We will measure and analyze this test setting difference as a covariate[27]. Group vs. individual administration will also be tested as a covariate[28]. We will not allow sites to collect their samples using paid data collection services, such as MTurk; as many CREP labs consist of student researchers who lack substantial financial resources[29]. The CREP would like to encourage students to collect data in a lab setting without incurring additional costs. However, two authors (Chartier and Hall) will collaborate on collecting one large (N = 500) pre-registered MTurk sample of US participants (with its own OSF) to compare the original sampling pool (i.e., MTurk) with the rest of the studied samples - which we will do by including a variable that specifies whether the sample is the MTurk sample or not in the planned multilevel models. We pre-registered and collected such a large MTurk sample size in order to have a sample that is sufficiently large (and thus likely has a small CI) and as close to the original sampling pool as possible to provide a more precise estimated comparison.

All participants will be asked whether they participated in this study before and will be excluded if they have (in part, to avoid "superturkers"). To further our ability to generalize beyond typical

[26] These materials were not always available for each student team. Thus, the data from each team was only included in the overall analyses if we were able to check raw data files and codebooks for accuracy and errors, which was only an issue for teams that collected data via Qualtrics. Given that we are using multilevel modeling, we only needed to check the quality of raw data files and codebooks for those data not collected via SoSciSurvey to be considered valid and usable. We automatically had access to the raw data and codebook for all samples collected via SoSciSurvey, so we always used those data rather than any data posted to the OSF. Analysis scripts, cleaned data, and narrative summaries were pedagogically important, but had no bearing on the results of the overall analyses reported in this paper.

[27] Given that most sites ended up moving to online collection due to the pandemic, most sites collected data online, and, thus, we opted to no longer use this variable.

[28] Again, the pandemic resulted in most sites testing participants individually and thus rendered this covariate unusable, so we dropped it from our analyses.

[29] We ended up allowing one site to use a paid data collection service to gather data for this study due to pedagogical reasons; however, this site was ultimately not included in analyses for other reasons.

university samples, we will also encourage (but not require) sites to collect an additional non-university sample ($N = 50$) by including a protocol for collecting non-university participants in their sampling plan - which will be rewarded with a higher author order and a CREP quality award. To track these efforts descriptively, we will measure which participants were recruited from the general public and which were recruited from a student body.[30]

Experimenters. Any trained undergraduate or graduate student researcher, research assistant, postdoctoral researcher, or faculty member can serve as the experimenter. Given the simplicity of the study design, no special expertise is required to conduct the study. During in-person testing, an experimenter should be unaware of the specific condition to which a participant is assigned (preferably via masking). We will only allow data collection via SoSciSurvey to streamline data collection and analyses.[31] The SoSciSurvey experiment code will be made publicly available, and Sophia Weissgerber will coordinate with translation teams to create experiment code for each site. Each site will be required to submit a video of their methodology for review by the CREP executive team (described below) and will then post to their site's OSF page.

Materials. We will use the same manipulations and outcome variable questions reported in Turri et al. (2015). We will also test two additional vignettes ("Fake Barn/Gerald" vignette from Colaço et al., 2014; "Diamond/Emma" vignette from Nagel, San Juan, & Mar, 2013) alongside the original Turri et al. (2015) Experiment 1 "Darrel" vignette (see Appendix B). The "Gerald" vignette had control conditions that were not exactly similar to Turri et al. (2015); therefore, we altered the conditions in order to be as closely matched to the "Squirrel/Darrel" vignette as possible.[32] Therefore, we altered the "Gerald" vignette and its controls to more closely resemble the "Darrel" vignette from Turri et al. (2015). We then pretested these vignettes for comprehension (about 90% comprehension rate across vignettes) and tested controls for expected rates (i.e., knowledge control viewed largely as knowledge, $M = 76.91$, $SD = 30.3$; ignorance control largely viewed as ignorance, $M = 10.12$, $SD = 21.61$; pretest means and standard deviations for each vignette reported in Appendix B). All materials used in this replication, including the details of these vignettes and related pretests, are available on our OSF page (Hall et al., 2018).

Each contributor site will pre-register their individual study on an OSF page connected to this parent pre-registration.[33] We will also record demographic information[1] that will include additional questions not reported in the original study for the use of exploratory analyses (e.g., participant race/ethnicity, years of education, age, geopolitical region of residence, geopolitical region of origin, and gender). In addition, the original study asked participant language proficiency by asking, "Did you take this test in your native language?" to exclude non-English speaking participants. However, given that many of our contributing sites are bi- or multilingual,

[30] While we originally planned to measure sample source, teams that opted to collect community samples did not add a question to the survey that would allow us to track which participants were from the community and which were typical university students. Thus, the only tracking of this methodological feature we implemented was asking labs whether they did any community sampling as part of their study after data collection.

[31] For pedagogical reasons, some teams were allowed to use Qualtrics surveys for their data collection.

[32] This sentence was edited slightly for clarity and accuracy.

[33] Some sites failed to preregister their study after being approved for data collection (or kept their OSF private).

we will instead ask how well participants speak the language in which they are being tested and if said language is their first language.[34]

Furthermore, each site will ask participants a set of funneled debriefing questions to assess participant knowledge of the study hypotheses (see Appendix B). To achieve this, each site will read their sample's responses one-by-one and exclude participants based on their level of awareness and note the particular reason for exclusion, and we will include these findings in an exploratory results section.[35] To support another project, we also have partnered with Satchell et al. (2018) to collect information about common participant study experiences using a short list of 12 questions (see Appendix C). Labs are encouraged, but not required to collect data from participants using this measure and will coordinate individually with Satchell et al. (2018). If labs participate, Satchell et al.'s questionnaire will only be inserted entirely at the end of our entire study package.[36]

Participant language. As one method of controlling for comprehension of the vignettes, participants will be asked how well they speak the language in which they were tested, using a 4-point scale ("very well", "well", "not very well", and "not well at all"). Teams for whom participants' primary language is other than English speakers must translate the study materials to their respective native language, and their translations must be approved by the PSA and CREP teams using the PSA procedures before they can be used with participants.

To be approved by the CREP team, translated materials for non-English speaking participants are asked to translate using the *Psychological Science Accelerator* (*PSA*) guidelines (https://psysciacc.org/translation-process/; Behling & Law, 2000; Moshontz et al., 2018). All study sites planning to test participants in the same target language will work together in a concerted, consolidated effort to translate study materials to the target language using these procedures, resulting in a unified translation that will be used by all same-language sites. To begin this process, materials will first be translated from English to the target language by "A" translators -- resulting in document Version "A" (i.e., forward translation). Version "A" will then be translated back from the target language to English by "B" translators independently -- resulting in Version "B" (i.e., backward translation). Both "A" and "B" translators must have knowledge of both English and the target language, have familiarity with both source and target cultures, and have experience in test development.[37] The "B" translators must be native English speakers and should not have worked with the specific test materials before. The backward translation and the original English test materials should be very similar.

[34] We did not ask participants if the language they were being tested in was their first language.

[35] Due to logistic issues, we opted to instead ask for bilingual volunteers to rate each response (three ratings per response) to determine if the participant met the exclusion criterion (coded as no = 0, maybe = 1, and yes = 2). If the sum of the three coders' ratings of a participant response was four or above, that participant was excluded from the overall analyses.

[36] Due to a miscommunication that occurred during the creation of the centralized survey on SoSciSurvey, the study experience questionnaire was included at the end of the core survey; thus, all collection sites ended up including this exploratory measure at the end of their survey. Data from this survey were not used in the present study but may be used for a future paper.

[37] Requiring translators to have experience with test development turned out not to be feasible; thus, we did not enforce this requirement.

Version “A” and “B” will then be discussed amongst translators “A” and “B” and the language coordinator, and discrepancies between version “A” and “B” will be identified and resolved among translators -- resulting in Version “C” (i.e., reconciled forward translation). Version “C” will then be tested on two non-academics fluent in the target language and then asked how they perceive and understand the translation. Possible misunderstandings are noted and again discussed as in the previous step. Finally, data collection labs read materials and identify any needed adjustments for their local participant sample. Adjustments are discussed with the language coordinator, who makes any necessary changes, resulting in the final version for each site. Final versions must then be submitted to the CREP for approval alongside their pre-registration, videotaped methods, and ethics approval.

Importantly, while using the above-described translation procedure, we will endeavor to ensure the equivalence across the original and translated versions. The established vignettes contain potentially unfamiliar nouns depending on participants’ cultural experiences (e.g., tornadoes do not occur in certain regions). Therefore, we will allow labs to substitute culturally specific nouns with locally relevant ones during the translation process described above (e.g., replace “tornado” with “typhoon”). Noun changes will be considered during the translation process as part of each translation team's effort to achieve equivalence in translations and will be noted on our OSF.

Data collection. Participants will be unaware of the specific hypotheses about Gettier intuitions and will not be informed that they are participating in a study about Gettier cases. Instead, participants will be told that this is a study about language using the exact language Turri et al. (2015) used in the original study (see Procedures)[38]. All participants will be randomly assigned (within each site) to one of three propositional knowledge conditions (i.e., knowledge, Gettier, or ignorance) and then counterbalanced within the three presented vignettes (six possible condition orders), always beginning with “Darrel” and then randomizing between “Emma” and “Gerald” (two possible vignette orders)[2]. Thus, approximately one-third of all participants will be randomly assigned to each belief condition in all three vignettes. Each participating lab is required to randomize using a predefined list of vignette/condition orders - which will be pre-programmed into the single survey software used to collect data (i.e., SoSciSurvey) at all sites. Although randomization will be pre-programmed into each site’s survey software, each site must describe the random assignment methods used in their pre-registered plan - which must be approved by the CREP executive team.

Procedure. Given that each contributing team must design their protocol using the standards and procedures set forth in this vetted manuscript, the details of each lab’s protocol will be consistent across labs. The CREP will only approve high-quality replication protocols that fit all the standards and procedures set forth in this manuscript. A typical procedural description would resemble the following.

Participants will first be given an Informed Consent form, which includes the following statement used by Turri et al. (2015): “There are no known risks to you for participating. We hope that our results will add to scientific knowledge about how language works.” Once they

[38] Some collection sites mistakenly worded the study purpose in their informed consent materials slightly differently, but none of these mistakes resulted in the experimenters giving away the true purpose of the study; thus, these sites were not excluded. More details for these deviations can be found on the OSF.

have provided their informed consent, participants will be presented with each of the three vignettes, randomly assigned and counterbalanced into a knowledge condition (to which the experimenter should be unaware via masking). The three vignettes will be presented in random order. Each vignette will be randomly assigned to a belief condition and counter-balanced so that each participant experiences all three vignettes ("Darrel", "Gerald", and "Emma") and all three belief conditions once (knowledge control, Gettier case, and ignorance control). Participants will be directed to their randomly assigned reading condition for each vignette (for full details of these vignettes, see Appendix B).

After participants have read each assigned vignette, they will then be asked to respond to several questions before moving on to the next vignette. As in Turri et al. (2015), participants will first respond to a knowledge attribution question followed by a comprehension question for exclusion purposes.[39] Then, participants will answer a question about whether it was reasonable or unreasonable for the protagonist to believe what they believed (Turri et al., 2015). For measuring these two dependent variables and the comprehension control variable, we will use the same procedure used in the original study (Turri et al., 2015). That is, participants will not be allowed to go back to a previous page and change their answer, and questions will always be asked in the same order (knowledge/ comprehension/ reasonableness) for each vignette.

After completing all confirmatory and exploratory questions for each vignette, participants will then be asked to answer a set of demographic, control, covariate, and study experience questions (see Appendix C)[40]. Control variables will include the language proficiency question described above and a set of funneled debriefing questions to check for explicit knowledge of our specific hypotheses[41]. Covariates include all demographic and other variables that are measured at each site by each participant, including the test setting (tested online vs. face-to-face; tested individually vs. in group, compensated vs. uncompensated), participant age, gender (men, women, other), and years of education. All other demographic questions will be reported for solely descriptive purposes. We will also collect a large swathe of site level variables (i.e., regional SES related information, local climate, crime prevalence, etc.) for the use of exploratory analyses. Also, as part of a Study Swap project (Chartier & McCarthy, 2018), contributing sites may opt into asking participants a set of additional questions about their study experience (Satchell, 2018; see Appendix C)[42]. We will collect responses to these questions, but we have no plans to use the information in any of our analyses.

Participating labs are free to compensate participants using the standards of their lab/university. This could include extra credit, research credit, money, gift cards, or no compensation (we will measure compensation as a covariate). However, as previously mentioned, we will not permit the

[39] This sentence originally described comprehension questions as control variables, but this description was inaccurate. Thus, we reworded the sentence to accurately reflect that comprehension questions were planned for exclusion purposes.

[40] Again, we reworded "control" variables to accurately reflect that we meant "exclusion" variables.

[41] Reworded "control" to "exclusion" for accuracy.

[42] As previously discussed, due to a miscommunication that occurred during the creation of the centralized survey on SoSciSurvey, the study experience questionnaire was included at the end of the core survey; thus, all collection sites ended up including this exploratory measure at the end of their survey. Data from this survey will not be used in the present study but may be used for a future paper.

use of online survey services where participants are paid (e.g., MTurk), except for one large MTurk sample that will be collected by two of the authors (Chartier and Hall).[43]

Data collection stopping rules and exclusions. Each site will pre-register a minimum target sample size of 50 as a part of their OSF pre-registration, which must be approved by the CREP executive team prior to data collection. To be approved, contributing labs must demonstrate a sufficient random assignment method and the ability to reach a minimum required sample size (after exclusions are accounted for). Contributors can stop collecting data when they meet their pre-registered target sample size, or when the overall data collection deadline passes. Overall data collection will be stopped when the April 1st, 2020, deadline passes, or once all contributors have reached their pre-registered target sample size.[44] Depending on the progress of the primary analyses, we cannot guarantee inclusion of projects submitted for review after this date.

Participants in any laboratory must be excluded for any one of the following reasons: (1) if the participant is not the majority age of their geopolitical region or older (unless parent/guardian waiver provided), (2) if the participant has taken part in a previous version of this study or in another contributor's replication of the same study, (3) if the participant fails to answer comprehension questions correctly, or (4) if the participant correctly and explicitly articulate knowledge of the specific hypotheses or specific conditions of this study when answering the funneled debriefing questions. We will also exclude participants who self-report their understanding of the tested language as "not well" or "not well at all". We based this exclusion criteria on a recent study that found that non-native English speakers who self-report as "very well" and "well" tend to score in the "intermediate" and "basic" categories on an English proficiency test respectively, while those who self-report as "not well" and "not at all" tend to score in the "below basic" category (Vickstrom, Shin, Collazo, & Bauman, 2015). All excluded data will be included in the data files on the overall OSF page, along with the particular reason for why they were excluded.

---

[1]Due to ethics considerations (e.g., EU policies regarding collecting certain demographic questions), individual sites may opt out of measuring specific descriptive demographic questions (e.g., race/ethnicity) on a case-by-case basis.

[2] Resulting in 6 propositional knowledge condition order combinations, and randomizing order of presenting vignettes (6 possible order combinations), resulting in 36 possible flows in SoSciSurvey.

Analysis Plan

Deviations from Approved Analysis Plan

Several deviations occurred in our analysis plan. First, the variables that we measured continuously had severe statistical violations that would have rendered our originally planned

[43] We ended up allowing one site to use a paid data collection site to gather data for this study due to pedagogical reasons; although, this site was ultimately not included in analyses for other reasons.

[44] The overall deadline to end data collection was extended to June 1st, 2021, to accommodate complications in data collection caused by the COVID-19 pandemic.

analyses uninterpretable and misleading. Thus, we converted those variables to binary (reasonableness, knowledge, and luck attributions). Because we converted these variables, many other deviations were required in order to produce the most valid and accurate analyses. In addition to these deviations, we also deviated from our plan regarding how we decided to build and test our models. These deviations occurred because the originally planned analyses did not follow common practices regarding multilevel modeling or were factually incorrect.

Reporting Site Level Analyses. Due to inconsistencies of student teams reporting their final site-level analyses, we do not have a full list of analyses that were conducted by student teams. Where we do have that information, it is reported above.

Converted Continuous Measures to Binary. Due to the extreme bimodal distributions, we observed in our data causing severe statistical violations that would have rendered our originally planned analyses uninterpretable, we had to convert the continuous measures to binary.

Model Testing. The null model was incorrectly described as including the primary predictor variable (condition). In practice, null models do not include any predictors. Thus, to ensure proper model testing, we did not follow this particular erroneous plan when constructing the true null model for our analyses. As per common practices for assessing multilevel models, we did not use the ICCs for random variables and instead relied on changes in the AIC and in the pseudo-R-squared values to determine how much variation in the data was accounted for by each random variable.

Because we converted our dependent variables to binary variables, it no longer made sense to assess whether judgements differed from chance based on their confidence intervals crossing the 50 mark. Given that we ended up assessing how much variation could be explained by vignettes using different model testing procedures, we opted to not test how correlated vignette responses were with each other. Given that the procedures described here for testing the random effects included in our planned multilevel models were erroneously described and do not follow current common practices in multilevel modeling, these procedures were not followed. Instead, current common practices were chosen before examining the data to ensure that conclusions derived from testing these models would be both statistically valid and uninfluenced by knowledge of how these changes might alter the final results.

The way these results were ultimately reported changed due to our decision to convert the continuous dependent variables to binary ones. Again, to ensure that we follow common practices in multilevel modeling, the variation accounted for by the random variables was assessed using changes to AIC and measures of pseudo-R-squared rather than the random slope variance. Because we did not ultimately calculate regression coefficients between vignettes, we did not utilize any family-wise error correction method.

We opted not to follow our plan to test covariate structures in order to better follow current common practices in multilevel modeling. Thus, we did not test which covariate structure fit the data best. Once again, to ensure that our analyses followed current common practices in multilevel modeling, we did not add each covariate to the model one at a time. Instead, all covariates were entered simultaneously to determine their influence before the focal predictor was entered into the model.

For reasonableness attribution, we altered our planned analyses in all of the same ways that we did for knowledge attribution. As previously discussed, we decided to make these changes to accommodate our conversion of the variable from continuous to binary and to ensure that we follow current common practices in multilevel modeling.

Proposed Analytic Strategy and Sample Size Justification

For this experimental mixed factorial design, we will analyze the primary and secondary hypothesized outcomes (i.e., knowledge and reasonableness attribution visual analogue scales, respectively) with multilevel modeling (for a visualization of this data structure, see Figure 3). In these analyses, participants in the contributing labs will be presented with a set of three stimuli (i.e., “Darrel”, “Gerald”, and “Emma” vignettes). As belief condition will also be random for each stimulus and each participant, this design feature will further give rise to a cross-classified data structure, where participants are nested within higher-level units formed by crossing two or more higher-level classifications with one another to fully account for the nesting of participants (i.e., participants are not only nested within their own labs, but also with regards to the conditions they have been exposed to).

The order of vignettes and their conditions will be randomized without replacement, such that participants will first be assigned to one of three vignettes (“Darrel”, “Emma”, or “Gerald”) in one of the three belief conditions (knowledge control, Gettier case, or ignorance control). The remaining two vignettes will then be presented in random order, each also randomized and counter-balanced to one of the remaining two belief conditions until all participants have been exposed to all three vignettes and all three conditions once. Participants will be asked several questions after reading each vignette. We will use this model to test whether the effects of the independent variable (i.e., knowledge condition) on the continuous dependent variables (i.e., visual analogue scale responses for knowledge and reasonableness) are robust to covariates/interactions (i.e., sensitivity test).[45]

[45] Due to the extreme bimodal distributions, we observed in our data causing severe statistical violations that would have rendered our originally planned analyses uninterpretable, we had to convert the continuous measures to binary.

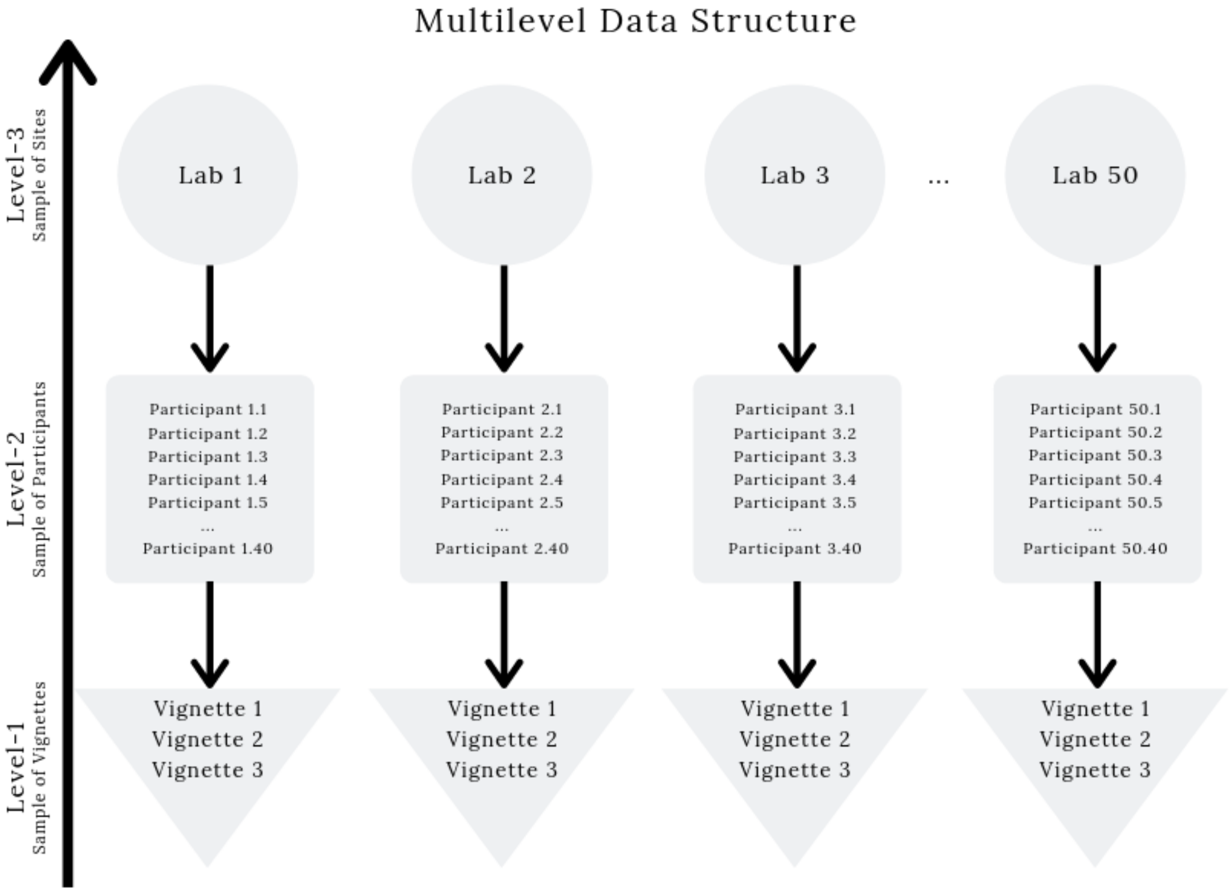


Figure 3: Data Structure. Total sample size includes a sample of labs (target lab sample size is 50) each nested with a sample of participants (minimum participant sample size is 50) which are cross classified with a sample of three vignettes (stimuli). Each assigned vignette is randomized to one of three conditions (ignorance control, Gettier case, or knowledge control).

Participant sample size. To estimate the required number of units needed in each level (i.e., vignettes, participants, labs) of our two primary three-level linear models (knowledge and reasonableness) for adequate power, we used R package “simr” (Green & MacLeod, 2016). We simulated 1,000 datasets (using the “powerSim” function) several times for different model specifications. We simulated distributions of the primary response variable based on the means and standard deviations of the data we collected during a pretest (Hall et al., 2018), which met assumptions for the analyses.

To estimate the difference in knowledge attribution rates between participants in the Gettier case condition and participants in the ignorance control condition for the power analysis, we used the Cramér’s *V* (.509) reported in Experiment 1 of Turri et al. (2015) and the observed unstandardized beta from our pretest data to roughly estimate a standardized fixed effect for the model (β = .5). We assumed that our test will likely find a smaller effect size closer to the average (i.e., regression toward the mean; β = .3) because our estimates were drawn from non-random samples using two imperfectly correlated measures and because an effect size of .5 is probably an extreme outlier within the distribution of all possible tests. We estimated a small difference (β = .10) in knowledge attribution rates between participants in the Gettier case condition and participants in the knowledge control condition based on our pretest data (Hall et al., 2018) and the small significant effects sometimes found in the literature, also assuming regression toward the mean for the same reason (e.g., Machery et al., 2017a; Starmans & Friedman, 2012).

We then explored several simulations with varying study parameters based on the pretest data we collected (Hall et al., 2018) and the original study (Turri et al., 2015). We investigated how this specified model could reach 90% power with an alpha of .05. We chose 90% power because we wanted to allow for a strong chance to detect a more accurate estimate of the effect sizes reported in the original publication, especially since there may be a small effect that went undetected in the original study (Hall et al., 2018). Additionally, effect sizes in the literature are often overestimates of the true effect size (Brandt et al., 2014; Greenwald, 1975; Open Science Collaboration, 2015; Simonsohn, 2013).

We used the R function "powerCurve" (Green & MacLeod, 2016) to simulate data along several participant site sample sizes while holding vignette sample size (*N* = 3) and lab sample size (*N* = 9) constant to determine what site sample sizes we need to achieve 90% power to detect a real (between-subjects) effect of condition on knowledge attribution. These simulations, available on our OSF project page (Hall et al., 2018), suggest that to be powered enough (90.2%, 95% *CI* [88.19, 91.97]) to detect a real between-subjects effect while accounting for the crossing and nesting of our data, we will need 3 vignettes per participant, 32 participants per lab, and 9 labs (288 total participants, 864 total observations; see Figure 4).

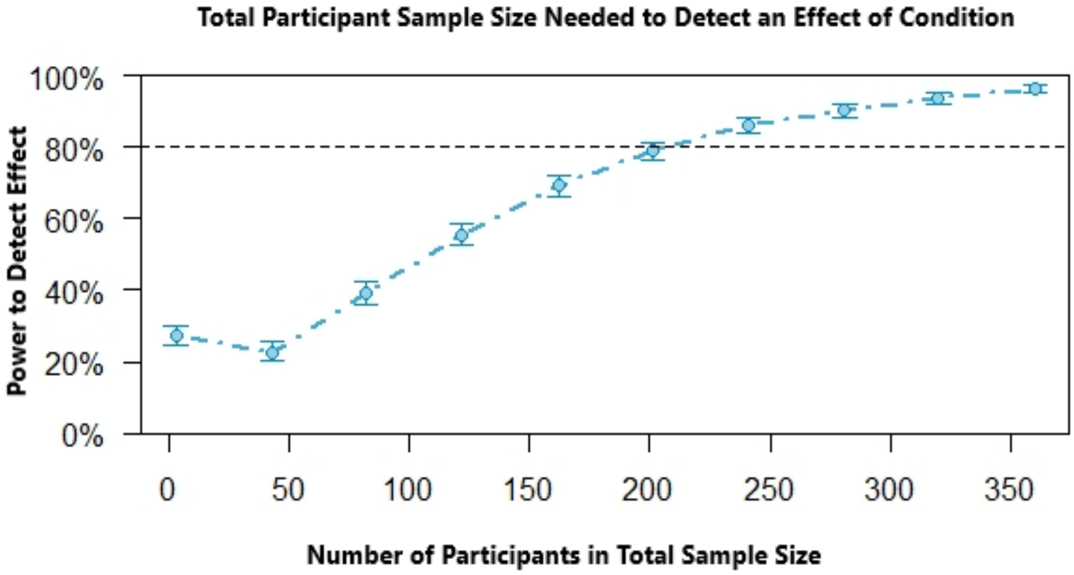


Figure 4: “powerCurve” (Green & MacLeod, 2016) plot for total number of participants needed (across all labs) to detect a small effect of condition (β = .1) on knowledge attribution, with vignette count (*N* = 3) and lab count (*N* = 9) held constant.

However, the power estimated by these conventional power analyses may differ non-trivially in the presence of effect heterogeneity - which has been shown to be an issue, even in large multilab studies (i.e., Many Labs) with minimal study variation (Kenny & Judd, 2019). For instance, when a study demonstrates some effect heterogeneity and a small to medium effect size, there is a non-trivial chance of finding a significant effect in the opposite direction from the average effect size reported in the literature as well as a non-trivial probability of detecting an effect in the wrong direction (i.e., the effect is positive, but the test actually shows a significant negative effect): This probability increases as N increases (Kenny & Judd, 2019). For these reasons, Kenny & Judd (2019) recently concluded that multiple smaller studies are preferable to a single large one, and that many smaller studies that vary those irrelevancies can likely tell us more than one single large study.

Rather than requiring a considerably larger participant sample size for each site in order to provide a better powered test to detect interaction effects of covariates, we instead weighed the important trade-offs between study feasibility for undergraduate students in a classroom setting and power for covariates. Due to the pedagogical focus of this project, we decided to prioritize study feasibility for undergraduate students (*N* = 50) rather than requiring a larger, more representative sample from each site (*N* > 100). Thus, any exploratory analyses of covariates and their interactions will be interpreted with caution.

Laboratory sample size. In terms of participating labs, we currently have 45 sites signed up. The PSA currently has over 450 labs within its global network, and the CREP currently works with 29 labs. Other multilab projects, such as ManyLabs, have similarly collected data from 20 to 30 contributing sites. In addition to this network of labs, AMMPS will make an additional call for contributing labs through APS.[46] Because of the CREP’s educational aims, we will continue to accept contributing labs until February 1, 2020, even after adequate power has been reached[47]. Given these numbers, past experiences of our team, and the low resource requirements of this study, we are confident in our ability to collect from at least 50 labs. This will give us more than adequate power to detect our primary effect of interest, as well as provide a rich and broad data set that other researchers can analyze to make secondary contributions.

Given that individual site samples may experience some data loss (we estimate at least 10% from comprehension exclusion based on pretest data, Hall et al., 2018), we will require a pre-registered minimum sample size of 50 participants (after exclusions) at each site to ensure that each data collection is reasonably powered. To incentivize sites to collect well powered samples and provide students with quality lab experiences, the CREP awards sites with a completion

[46] Due to being overcapacity for onboarding new labs and concerns about our small administrative team being able to manage more collection sites, we opted to not make an additional call for contributors.

[47] Due to the deadline for stopping data collection being extended through to June 1st, 2021, because of the COVID-19 pandemic, we continued accepting labs until April of 2021.

certification award for meeting the required sample size. To qualify for the CREP completion award for this study, a site must sample at least 50 participants. The CREP completion award is a certificate presented to participating lab members for their high-quality work upon completion of the CREP study.

Stimulus sample size. To determine which vignettes (i.e., stimuli) to sample from the experimental philosophy literature, we first searched the literature thoroughly for all articles relating to Gettier intuitions. We then evaluated the vignettes found in these articles based on several criteria: similarity, quality, and influence. Because our goal is to replicate findings from Turri et al. Experiment 1 (which used a counterfeit-object Gettier case), we decided to only sample other counterfeit-object vignettes to test the generality of this class of Gettier-type cases, in lieu of testing Gettier-type cases more broadly. Thus, we first determined if a given vignette was a counterfeit-object type Gettier case or if it was a different type of Gettier case (e.g., evidence replacement), and then kept only counterfeit cases. Then, we noted whether a given vignette had matched controls similar to those found in Turri et al. (2015) and kept only those that did. We then evaluated the influence of the remaining vignettes based on how many times an iteration of the vignette has been tested in the literature. Through this process, the "Gerald" vignette (i.e., fake barn case) and the "Emma" vignette (i.e., the counterfeit diamond case) were selected. Of note, the "Gerald" barn case vignette was primarily chosen for its influence in the literature - as its control conditions were not perfectly matched with the "Squirrel/Darrel" vignette conditions and had to be altered to be used.[48]

Planned Analyses

In total, [X] labs applied to participate in this multilab replication. [X] labs were unable to participate, [X] did not collect enough data; [X] dropped out prior to data collection, resulting in a final lab count of [X]. Contributing labs represent [X] continents ([X from Africa, X from South America, X from North America, X from Asia, X from Europe, and X from Oceania) with participants residing in [X] geopolitical regions [X from Brazil, X from Switzerland, X from Singapore, and so on]. [X labs committed to collecting the minimum participant sample size (N = 40), and X labs committed to collecting a larger, more representative sample (N = 100) for the purposes of exploratory analyses. All participating labs submitted their dataset and analysis report for review to the CREP team. All datasets were required to be submitted using a template dataset that must pass a quality check (raw data files checked for errors; post analysis scripts, codebook, and cleaned data files checked for errors; and narrative summary of project findings compared to data and analysis for errors)[49]. For strictly educational purposes, contributors chose

[48] This sentence was edited slightly for clarity and accuracy. We previously stated that the Gerald vignette was chosen, in part, because it had minimally matched controls available; however, this was inaccurate. The controls used in the study from which it was taken were not similar to those used in the Turri et al. (2015) study and ultimately had to be altered to be usable. Thus, the primary reason for its inclusion was its prevalence in the epistemics literature.

[49] These materials were not always available for each student team. Thus, the data from each team was only included in the overall analyses if we were able to check raw data files and codebooks for accuracy and errors, which was only an issue for teams that collected data via Qualtrics. Given that we are using multilevel modeling, we only needed to check the quality of raw data files and codebooks for those data not collected via SoSciSurvey to be considered valid and usable. We automatically had access to the raw data and codebook for all samples collected via SoSciSurvey, so we always used those data rather than any data posted to the OSF. Analysis scripts, cleaned data,

which analyses to perform on the effects of condition on the continuous knowledge attribution variable (Y_1) and the continuous reasonableness attribution variable (Y_2) on their site sample.[50] We did not provide any specific plans for sites to analyze their data, and instead allowed sites to choose which analyses to perform (Silberzahn et al., 2018). Full details of these analyses are available via this study's pre-registration on the OSF project page (https://osf.io/n5b3w/).

Although, we did not direct instructors and students to use specific analyses, we did provide support as they determined which analyses to pre-register and provided feedback on the subsequent analysis reports at each site[1], [X sites chose to perform a mixed effect ANOVA; X sites chose to perform a two-level linear regression analysis; X sites dichotomized the visual analogue scale responses and performed a two-level logistic regression analysis, and so on][51]. We also provided a data template with variable naming conventions on our OSF page which contributor sites were required to use when submitting their sample data (available on our OSF).[52] The results of site level analyses will be included on each site's pre-registered OSF page. The datasets from each lab were included, regardless of their results, providing a more unbiased study of the effect.

The typical goal of an RRR is to provide a more precise effect size estimate by combining the results of a number of independently conducted direct replications - typically using a meta-analytic approach. Our goal for this RRR continues this trend; however, we instead aggregated individual participant data from each site in order to conduct a pair of multilevel linear regression analyses that account for the nesting of data and treats the tested vignettes as a random factor. The purpose of these analyses is to determine a more accurate effect size estimate for Gettier intuitions (Brandt et al., 2014), rather than to "fail" or "succeed" at replicating the original results. Therefore, we combined all site data that passed the CREP quality check (see above) into one data file containing all individual participant data [X were excluded due to DESCRIBE QUALITY ISSUES; X labs remain in the primary analyses], which we analyzed with two multilevel models: one on the continuous knowledge attribution measure and one on the continuous reasonableness attribution measure[53]. We performed these analyses to test whether the effects of the primary between-subjects factor (belief condition and laboratory) and the exploratory within-subjects factor (vignette condition) on the given outcome variable (knowledge or reasonableness) are robust to covariates/interactions (i.e., sensitivity test).

Authors, Jordan Wagge and Braeden Hall, wrote the R scripts, simulated data, and analyzed power for the overall and site analyses before any data were collected. The two multilevel linear regression analysis R scripts include assumption tests and analyses of the overall effect of belief

and narrative summaries were pedagogically important, but had no bearing on the results of the overall analyses reported in this paper.

[50] The continuously measured variables were ultimately converted to binary measures due to severe statistical violations caused by highly bimodal distributions that would have rendered our planned analyses uninterpretable.

[51] Due to inconsistencies of student teams reporting their final site-level analyses, we have opted not to report which analyses were conducted by student teams.

[52] This data template was ultimately not used due to the vast majority of samples being collected via the centralized data collection via SoSciSurvey.

[53] As previously discussed, the continuous dependent variables were ultimately converted to binary variables to ensure the statistical conclusion validity of our planned analyses.

condition (Knowledge, Gettier, Ignorance) on the primary outcome (knowledge attribution) and the secondary outcome (reasonableness attribution). Within these models, vignette was tested as a random (within-subjects) factor, condition was tested as a fixed (between-subjects) factor, and labs were tested as a random (between-subjects) factor. We also fitted these models with several exploratory covariates, including participant gender, years of education, age, and three test setting lab variables (online vs. in person; in group vs. individually; compensated or not compensated).[54] This will allow us to look at the extent to which the use of Gettier intuitions are prevalent within this sample of the general public. Exploratory analyses will also allow us to test the extent to which there are individual, lab, and stimulus differences; although, we will be cautious when interpreting these results. Other exploratory analyses will include the other covariates described below that are collected at every site.

Before we performed these analyses, we tested assumptions on our data. We first checked the data for linearity. [If a non-random trend emerges, we will then attempt to include a higher order (geopolitical region level-4 units) to see if that resolves the issue. This will suspend all power considerations reported earlier in the manuscript] For these two multilevel linear regression analyses, level-1 units (vignettes) were tested as a random factor crossed with the level-2 units (participants) that are nested in the level-3 units (lab sites). [If we have enough participating geopolitical regions (>20 geopolitical regions) to provide adequate power and if our model ends up requiring adding another higher order to correct for data dependence, we will then test whether adding geopolitical region of residence as a level-4 cluster unit, grouped into UN regions (i.e., Africa, Asia, North America, Oceania, etc.) improves the model or not.]

Knowledge attribution. Given that we are primarily interested in the relationship between the hypothesized level-2 between-subjects predictor (X_1) and the two hypothesized outcome variables (knowledge, Y_1; and reasonableness, Y_2), we first performed the analysis using solely the primary hypothesized independent variable (belief condition) without any other covariates for the purpose of trying to estimate the overall individual level effect (fixed slope) on the primary hypothesized outcome (i.e., null model). In other terms, we determined the effect on knowledge attribution across all samples, not accounting for covariates, vignette differences, or lab differences.[55] We found that the overall effect of belief condition was [insignificant/small/medium/large, β = .XX, 95% CIs [X.XX, X.XX]]. We then tested the model fit for each analysis using likelihood ratio (LR) chi-square difference tests to determine whether each unit level should be tested as a random or fixed factor and whether covariates improved the model (Gelman & Hill, 2007, Chapter 17; see Table 1).

To assess the model fit of our data, we used the commonly used nested model test using maximum likelihood estimation (Snijders & Bosker, 2012). Next, we wanted to determine if the effects of the primary independent variable (belief condition) on knowledge attribution differed by vignette, participant, or lab. To accomplish this, we built an unconditional base model for the knowledge attribution predictor to calculate the intra-class correlation coefficients (ICC) for

[54] Again, we opted to not use two of the three test setting covariates (online vs in person; individually vs in group) due to the COVID-19 pandemic causing most teams to collect data online.

[55] This null model was incorrectly described as including the primary predictor variable (condition). In practice, null models do not include any predictors. Thus, to ensure proper model testing, we did not follow this particular erroneous plan when constructing the true null model for our analyses.

vignette, participant, and lab variation. The ICCs for vignettes, participants, and labs in the dataset measures the percentage of variation explained by each level, such that vignettes accounted for [X.XX%, 95% CI [X.XX, X.XX] of the raw variation in the dataset, participants accounted for [X.XX%, 95% CI [X.XX, X.XX] of the raw variation in the dataset, and labs accounted for [X.XX%, 95% CI [X.XX, X.XX] of the raw variation in the dataset.[56]

Given that this base model did not include any other predictor variables, the total effect on knowledge attribution for a typical vignette within a typical participant corresponds directly with the fixed slope [X]; such that participants in the Gettier condition attributed knowledge across a visual analogue scale X more/less than participants in the knowledge control condition, and X more/less than participants in the ignorance control condition (see Figure 5).[57] To calculate the overall effect on knowledge attribution, we first calculated the given effect of each vignette - which we then used to calculate the random intercept variance [X]. [Because the CIs for knowledge attribution in the [knowledge control/Gettier case/ignorance control condition] [do/do not] cross 50, we [can/cannot] conclude that participants' judgments differed from chance.][58]

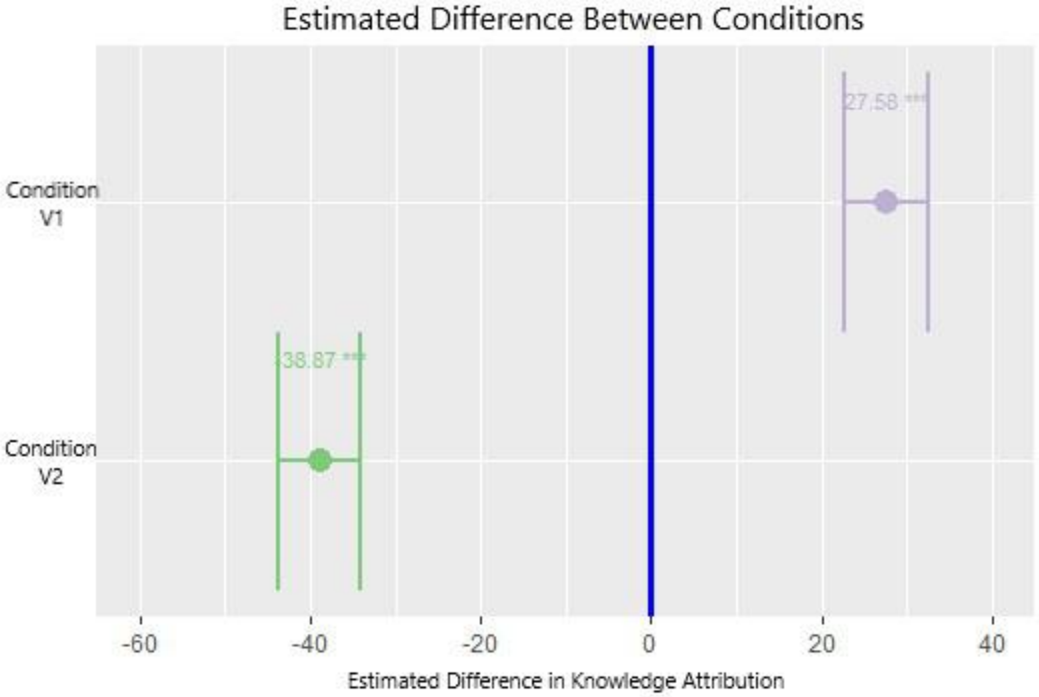


Figure 5. Example plot using simulation data to visualize the predicted difference between each condition, where Condition V1 is the estimated predicted difference between the Gettier case and the knowledge control (b = X.XX, t(XXX) = X.XX, p = .XX, 95% CI [X.XX, X.XX]) and

[56] As per common practice for assessing multilevel models, we did not use the ICCs for random variables and instead relied on changes in the AIC and in the pseudo-R-squares to determine how much variation in the data was accounted for by each random variable.

[57] Mislabeled as Figure 4.

[58] Because we converted our dependent variables to binary variables, it no longer made sense to assess whether judgements differed from chance based on their confidence intervals crossing the 50 mark.

Condition V2 is the estimated predicted difference between the Gettier case and the ignorance control (*b* = X.XX, *t*(XXX) = X.XX, *p* = .XX, 95% CI [X.XX, X.XX]).

Next, the level-1 residual [X] corresponds to the deviation of the specific effects of attributing knowledge within a given vignette from the overall effect of attributing knowledge across all vignettes - demonstrating that the intercept [varies/does not vary]. Given that the subsequent random intercept variance [X], was [small-large], this indicates that individual participants have [more/the same] opportunities of attributing knowledge in some vignettes than in others. [indicate which vignettes were likely to result in more/less knowledge attribution in which condition]. [None/Two/Three] of the sampled vignettes were significantly correlated to each other: a set of Pearson correlation coefficient tests indicated that there was a [non-/small/medium/large] significant positive/negative association between the knowledge attribution response rates in the Darrel vignette and the knowledge attribution response rates in the Emma vignette, (r(XXX) = .XX, p = .XXX), a [non-/small/medium/large] significant positive/negative association between the Darrel vignette and the Gerald vignette, (r(XXX) = .XX, p = .XXX), and a [non-/small/medium/large] significant positive/negative association between the Emma vignette and the Gerald vignette, (r(XXX) = .XX, p = .XXX). These [small/medium/large] [non-/significant] correlations coupled with the [low/moderate/high] Intraclass Correlation Coefficient that suggests that the vignette factor accounted for X.XX%, 95% CI [X.XX, X.XX] of the raw variation in the dataset provides [weak/moderate/mixed/strong] evidence that [none/at least two/all three] of our repeated measures (vignettes) demonstrated [poor/fair/excellent] reliability with each other. [When interpreting these ICCs, we will use Rosner's (2006) suggested criteria, where an ICC of less than 0.4 indicates poor reliability, an ICC greater than or equal to 0.4 but less than 0.75 indicates fair to good reliability, and an ICC great than or equal to 0.75 indicates excellent reliability.][59]

Table 1: Multilevel models of knowledge attribution;

(Dependent variable: Knowledge attribution;

Fixed: intercept, belief condition (base = Gettier))

	Constant		Knowledge		Ignorance	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.

[59] Given that we ended up assessing how much variation could be explained by vignettes using different model testing procedures, we opted to not test how correlated vignette responses were with each other.

I = Null (includes condition)	X.X	X.X	X.X	X.X	X.X	X.X
II = I + vignette	X.X	X.X	X.X	X.X	X.X	X.X
III = II + lab	X.X	X.X	X.X	X.X	X.X	X.X
IV = III + test setting	X.X	X.X	X.X	X.X	X.X	X.X
V = IV + education	X.X	X.X	X.X	X.X	X.X	X.X
VI = V + gender	X.X	X.X	X.X	X.X	X.X	X.X

Fixed (Continued): Vignette (base = Darrel vignette)

	Gerald		Emma	
Model	Est.	s.e.	Est.	s.e.
I	X.X	X.X	X.X	X.X

II	X.X	X.X	X.X	X.X
III	X.X	X.X	X.X	X.X
IV	X.X	X.X	X.X	X.X
V	X.X	X.X	X.X	X.X
VI	X.X	X.X	X.X	X.X

Fixed (Continued): Test setting (base = in person; base = individually; base = not translated)

	Online		In group		Translated	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
I	X.X	X.X	X.X	X.X	X.X	X.X
II	X.X	X.X	X.X	X.X	X.X	X.X
III	X.X	X.X	X.X	X.X	X.X	X.X
IV	X.X	X.X	X.X	X.X	X.X	X.X
V	X.X	X.X	X.X	X.X	X.X	X.X

VI	X.X	X.X	X.X	X.X	X.X	X.X

Fixed (Continued): Geopolitical Region (base = U.S.A.)

	Turkey		Brazil		China		And, so on...
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.	...
I	X.X	X.X	X.X	X.X	X.X	X.X	...
II	X.X	X.X	X.X	X.X	X.X	X.X	...
III	X.X	X.X	X.X	X.X	X.X	X.X	...
IV	X.X	X.X	X.X	X.X	X.X	X.X	...
V	X.X	X.X	X.X	X.X	X.X	X.X	...
VI	X.X	X.X	X.X	X.X	X.X	X.X	...

Fixed (Continued): Gender (base = female)

	Male	
Model	Est.	s.e.

I	X.X	X.X
II	X.X	X.X
III	X.X	X.X
IV	X.X	X.X
V	X.X	X.X
VI	X.X	X.X

Random

	Vignette			Participant			Lab			Log-Lik
Model	Var.	s.e.	s.d.	Var.	s.e.	s.d.	Var.	s.e.	s.d.	
I	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
II	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
III	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
IV	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X

V	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
VI	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X

With the fixed-effects and random-effects specified, we then added in explanatory variables. In this phase, we wanted to test how knowledge attribution rates differ across vignettes and labs when also accounting for all predictors. That is, we wanted to know whether vignette and lab factors account for the random slope. For this purpose, we built a constrained and augmented intermediate model by adding level-2 predictors (belief condition, gender, age, and education), level-3 predictors (in group vs. individually and online vs. in person), and their cross-level interactions, and then performed a likelihood ratio test for the given outcome variable to determine whether considering the cluster-based (vignettes, participants, and labs) variation of the effect of the lower level variables improves the model fit ($X(1)$ = X.XX, p = .XX). The results were [non-significant/significant], suggesting that addition of the random slopes [did/did not] improve the fit of the model. Therefore, the [fixed/random-intercept/slope model] appears to be the best fit.[60]

In the last phase of this analysis, we created a final model based on our prior models by either including the random terms or not for each factor, and then we added the cross-level interactions for knowledge attribution. By doing this, we can infer whether the effects of the independent variables on the dependent variable are robust to covariates/interactions (i.e., sensitivity test). In terms of the level-2 effect, the first three models provide us with two terms of interest, the fixed slope [X] and the random slope variance [X] for each level. The fixed slope represents the general effect of the primary independent variable (belief condition) on knowledge attribution. Condition [did/did not] significantly predict knowledge attribution rates (B = XX.XX, β = .XX, p =.XXX) and [significantly accounted for X.XX% of the variance/did not significantly account for any of the variance], (R^2 = .XX, F(X,XX) = X.XX, p = .XX).[61]

The residual term associated with the primary independent variable [X] provides a yardstick for determining the size of the effect variation and corresponds to the deviation of the specific effects of the primary independent variable across all vignettes and laboratories (Sommet & Morselli, 2017). The random slope variance for vignettes was [X, p = .XX], indicating that the variation of the effect of the primary independent variable (belief condition) from one vignette to another was [small/moderate/large/non-significant]. The random slope variance for labs was [X,

[60] Given that the procedures described here for testing the random effects included in our planned multilevel models were erroneously described and do not follow current common practices in multilevel modeling, these procedures were not followed. Instead, current common practices were chosen prior to implementing the analyses to ensure that conclusions derived from testing these models would be both statistically valid and uninfluenced by knowledge of how these changes might alter the final results.

[61] The way these results were ultimately reported changed due to our decision to convert the continuous dependent variables to binary ones.

p = .XX], indicating that the variation of the effect of the primary independent variable (belief condition) from one lab to another was [small/moderate/large/nonsignificant] (see Figure 5 for visualization of lab variation).[62]

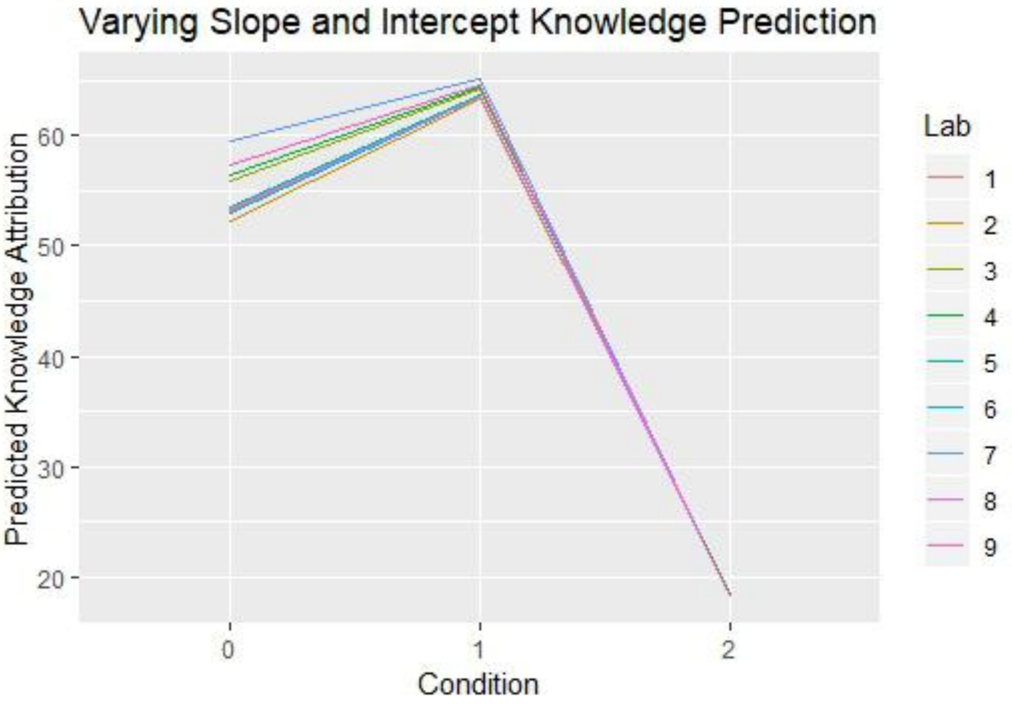


Figure 5. Example plot using simulation data to visualize the knowledge attribution multilevel model that allows for a random intercept and random slope for labs, where 0 = Gettier cases, 1 = knowledge controls, and 2 = ignorance controls. Actual data will be plotted based on the model of best fit.

False discovery rate. To correct for family-wise error rates that arise from testing related dependent variables, we will use a corrected alpha cut-off criterion. In our MTurk pretest of US participants, the knowledge variable and the exploratory ability/luck variable were significantly correlated in each vignette (“Gerald”, *r* = .476, *p* < .001; “Emma”, *r* = .434, *p* < .001; “Darrel”, *r* = .592, *p* < .001). Whereas the knowledge dependent variable and reasonableness dependent variable were weakly significantly correlated in two of the vignettes (“Emma”, *r* = .202, *p* = .009; and “Darrel”, *r* = .323, *p* < .001), and non-significant in the third (“Gerald”, *r* = .128, *p* = .099). These preliminary data demonstrate a need to lower our false discovery rate to correct for the family-wise error rate of three related tests which we will do by using the Benjamini–

[62] Again, to ensure that we follow common practices in multilevel modeling, the variation accounted for by the random variables was assessed using changes to AIC and measures of pseudo-R-squared rather than the random slope variance.

Hochberg procedure as well as performing bootstrapping to obtain confidence intervals for each correlation using Fisher's transformation.[63]

Covariate analysis plan. We then fit a covariance structure to this final model that specified the form of the variance-covariance matrix. We attempted fitting data with three common structures (variance components, diagonal, and unstructured) and tested differences between these fits with a goodness of fit test (BIC) to determine which covariance structure fits the data best (see Table 2); [results suggest that an unstructured covariance structure fits the data best, $X(1)$ = X.XX, BIC = XX.XX, p = .XXX.[64] If the model has convergence problems, we will try to increase the number of iterations, change tolerance levels, change optimization methods (e.g., BOBYQA optimizer instead of the Nelder-Mead optimization routine), and simplify the model by removing the random effect of vignette and the random effect of lab, in that order.

Table 2: *Covariance Structure*

Covariance Structure	(X)(1)	BIC	*p*
variance components	X.XX	X	.XX
diagonal	X.XX	X	.XX
unstructured	X.XX	X	.XX

In a covariate analysis, we then added each covariate to the model and compare the models to determine whether each covariate improved the model or not (see Table 3 for model comparisons; see Table 4 for beta coefficient estimates for each predictor).[65] However, because most of our site samples likely lacked adequate power to detect the effects of covariates and were

[63] Because we did not ultimately calculate regression coefficients between vignettes, we did not utilize any family-wise error correction method.

[64] Again, we opted not to follow this plan in order to better follow current common practices in multilevel modeling. Thus, we did not test which covariate structure fit the data best.

[65] Once again, to ensure that our analyses followed current common practices in multilevel modeling, we did not add each covariate to the model one at a time. Instead, all covariates were entered simultaneously to determine their influence.

not very representative or balanced regarding participant-level covariates, results from Table 3 and 4 should be interpreted carefully.

Table 3: Covariates – Model Comparisons

Covariates	(X)(1)	BIC	*p*
Participant Covariates			
Years of education	X.XX	X	.XX
Age	X.XX	X	.XX
Gender	X.XX	X	.XX
Lab Covariates			
Online vs. In person	X.XX	X	.XX
Individual vs. In group	X.XX	X	.XX
Compensated vs. Not	X.XX	X	.XX

Table 4: Covariates – Unstandardized (*B*) and Standardized (*β*) Beta Coefficients

Covariate	*B*	*SE B*	β	*t*	*p*

Participant Covariates					
Years of education	XX.XX	X.XX	.XX	X.XX	.XX
Age	XX.XX	X.XX	.XX	X.XX	.XX
Gender	XX.XX	X.XX	.XX	X.XX	.XX
Lab Covariates					
Online vs. In person	XX.XX	X.XX	.XX	X.XX	.XX
Individual vs. In group	XX.XX	X.XX	.XX	X.XX	.XX
Compensated vs. Not	XX.XX	X.XX	.XX	X.XX	.XX

Reasonableness attribution. We then analyzed an identical multilevel model for the reasonableness attribution dependent variable. We found [no/ a small/medium/large] effect of condition on reasonableness attribution (see Table 5), indicating that differences in knowledge attribution rates [are/are not] due to perceived differences of what is reasonable for a given protagonist to believe. Condition [did not] significantly predicted reasonableness attribution rates (*B* = XX.XX, *β* = .XX, *p* =.XXX) and [did not] significantly account[ed] for [any/X.XX%] of the variance, (R^2 = .XX, *F*(X,XX) = X.XX, *p* = .XX). The residual term associated with the primary independent variable [X] provides a yardstick for determining the size of the effect variation and corresponds to the deviation of the specific effects of the primary independent variable across all vignettes and laboratories (Sommet & Morselli, 2017). The random slope variance for vignettes was [X, *p* = .XX], indicating that the variation of the effect of the primary independent variable (belief condition) from one vignette to another was [small/moderate/large/non-significant]. The random slope variance for labs was [X, *p* = .XX], indicating that the variation of the effect of the primary independent variable (belief condition) from one lab to another was [small/moderate/large/nonsignificant] (see Figure 4 for visualization of lab variation).[66]

Table 5: Multilevel models reasonableness attribution;

(Dependent variable: Reasonableness attribution;

Fixed: intercept, belief condition (base = knowledge case))

	Constant		Gettier		Ignorance	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
I = Null (includes condition)	X.X	X.X	X.X	X.X	X.X	X.X
II = I + vignette	X.X	X.X	X.X	X.X	X.X	X.X
III = II + lab	X.X	X.X	X.X	X.X	X.X	X.X
IV = III + test setting	X.X	X.X	X.X	X.X	X.X	X.X
V = IV + education	X.X	X.X	X.X	X.X	X.X	X.X
VI = V + gender	X.X	X.X	X.X	X.X	X.X	X.X

[66] For reasonableness attribution, we altered our planned analyses in all of the same ways that we did for knowledge attribution. As previously discussed, we decided to make these changes to accommodate our conversion of the variable from continuous to binary and to ensure that we follow current common practices in multilevel modeling.

Fixed (Continued): Vignette (base = “Darrel” vignette)

	Gerald		Emma	
Model	Est.	s.e.	Est.	s.e.
I	X.X	X.X	X.X	X.X
II	X.X	X.X	X.X	X.X
III	X.X	X.X	X.X	X.X
IV	X.X	X.X	X.X	X.X
V	X.X	X.X	X.X	X.X
VI	X.X	X.X	X.X	X.X

Fixed (Continued): Test setting (base = in person; base = individually; base = compensated)

	Online		In group		Not Compensated	
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.
I	X.X	X.X	X.X	X.X	X.X	X.X

II	X.X	X.X	X.X	X.X	X.X	X.X
III	X.X	X.X	X.X	X.X	X.X	X.X
IV	X.X	X.X	X.X	X.X	X.X	X.X
V	X.X	X.X	X.X	X.X	X.X	X.X
VI	X.X	X.X	X.X	X.X	X.X	X.X

Fixed (Continued): Geopolitical Region (base = U.S.A.)

	Turkey		Brazil		China		And, so on...
Model	Est.	s.e.	Est.	s.e.	Est.	s.e.	...
I	X.X	X.X	X.X	X.X	X.X	X.X	...
II	X.X	X.X	X.X	X.X	X.X	X.X	...
III	X.X	X.X	X.X	X.X	X.X	X.X	...
IV	X.X	X.X	X.X	X.X	X.X	X.X	...
V	X.X	X.X	X.X	X.X	X.X	X.X	...

VI	X.X	X.X	X.X	X.X	X.X	X.X	...

Fixed (Continued): Gender (base = female)

	Male	
Model	Est.	s.e.
I	X.X	X.X
II	X.X	X.X
III	X.X	X.X
IV	X.X	X.X
V	X.X	X.X
VI	X.X	X.X

Random

	Vignette			Participant			Laboratory			Log-Lik

Model	Var.	s.e.	s.d.	Var.	s.e.	s.d.	Var.	s.e.	s.d.	
I	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
II	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
III	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
IV	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
V	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X
VI	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X	X.X

Exploratory analysis plan. One unique facet of CREP studies is that student contributors are encouraged to add extensions to their replication (referred to as Direct+Plus replications), which could involve testing conditions or measures in their local sample after testing the direct portion of the replication protocol (retaining all direct aspects put forward in this protocol). Attempting to replicate prior research provides students with experience in methodology and research design and encouraging students to design and test their own extensions provides them experience in planning, pre-registering, and testing their own hypotheses. Each contributing site that elects to participate in a Direct+Plus replication will be required to pre-register an independent power analysis and required sample size (for adequate power) that includes all planned Direct+Plus analyses. We will provide assistance with these power analyses by helping simulate data in R. All teams participating in the same Direct+Plus extension may pool samples to meet the required sample size.

Although we will require a relatively small sample size for the purpose of the confirmatory analyses described above (the primary stream of data collection), we will also facilitate a secondary stream of data collection by rewarding contributors who commit to collecting a larger, more representative sample ($N = 100$) with recognition through author order as well as a CREP quality award. Data from this exploratory stream of data collection will be released as part of the data release plan to allow other researchers to perform exploratory analyses on this larger set of data. We will report these analyses in an exploratory section in the post-data Phase 2 manuscript as well as on our OSF page.

[1] We may write about these analytic choices in a later publication.

Appendix B

Vignettes

All translated versions can be found here: https://osf.io/rnzqm/

Fake Barn/Gerald Vignette

For the "Fake Barn/Gerald" vignette, all participants first read:

> 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 new houses in the place of the destroyed houses. These new houses were rebuilt with all the materials necessary for them to look exactly like the original houses from the road, and they are also fully furnished and can now be used as actual housing."

- In the **knowledge** condition, participants then read:

  > Having just entered the tornado-ravaged area, Gerald notices the many houses lining the roads. When he tells Andrew 'That's a house,' the object he sees and points at is a real house that has survived the tornado and not one of the new houses.

- In the **ignorance** condition, participants then read:

  > Having driven through the tornado-ravaged area, Gerald has encountered many of these fake houses. When he tells Andrew 'That's a house,' the object he sees and points at is a fake house that was built after the tornado and is not actually a house.

- In the **Gettier** condition, participants then read:

  > Having just entered the tornado-ravaged area, Gerald has not yet encountered any fake houses. When he tells Andrew 'That's a house,' the object he sees and points at is a real house that has survived the tornado and not one of the fake houses.

Measured Variables for Fake Barn/Gerald

Presented in this order:

- Primary knowledge probe (from Turri et al., 2015):
  - **"Gerald ______ that he is pointing at a real house."**
    - Visual analogue scale, 0-100:
      - [only believes <-------------------------> knows]
- Comprehension question (from Turri et al., 2015):

    - **“Gerald is pointing at a _____ house. ”**
        - Binary: [real/fake]
- Reasonableness probe (from Turri et al., 2015):
    - **“It is _____ for Gerald to think that he is pointing at a real house.”**
        - Visual analogue scale, 0-100:
            - [unreasonable <------------------------> reasonable]
- Luck/Ability probe (from Turri, 2016b)
    - **“Gerald got the ______ answer because of his _____.”**
        - Requires two responses:
            - Binary: [right/wrong]
            - Visual analogue scale, 0-100:
                - [ability/inability <------------------------> (good luck/bad luck]
- Alternative knowledge probe (from Nagel et al., 2013)
    - **“In your view, which of the following sentences better describes Gerald’s situation?”**
        - Binary: ["Gerald knows that the house he is pointing at is a real house." OR "Gerald feels like he knows that the house he is pointing at is a real house, but he doesn't actually know that it is."]

Diamond/Emma Vignette

For the Emma vignette, all participants first read:

> Emma is shopping for jewelry. She goes into a nice-looking store and selects a 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 the **knowledge** condition, participants then read:

> However, this particular store has very honest employees who have a really positive reputation for their guaranteed real diamonds; in the tray Emma chose, all of the pendants had real diamonds rather than fake cubic zirconium stones (and the one she chose was really nice).

- In the **ignorance** condition, participants then read:

> Unfortunately, this particular store has very dishonest employees who have 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 (and the one she chose was in fact fake).

- In the **Gettier** condition, participants then read:

> Unfortunately, this particular store has very dishonest employees who have 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).

Measured Variables for Diamond/Emma

Presented in this order:

- Primary knowledge probe (from Turri et al., 2015):
  - **"Emma ______ that she chose a necklace made of diamonds."**
    - Visual analogue scale, 0-100:
      - [only believes <------------------------> knows]
- Comprehension question (from Turri et al., 2015):
  - **"Emma chose a necklace made of _____ ."**
    - Binary: [cubic zirconium stones/diamonds]
- Reasonableness probe (from Turri et al., 2015):
  - **"It is _____ for Emma to think that she chose a necklace made of diamonds."**
    - Visual analogue scale, 0-100:
      - [unreasonable <------------------------> reasonable]
- Luck/Ability probe (from Turri, 2016b)
  - **"Emma got the ______ answer because of her _____."**
    - Requires two responses:
      - Binary: [right/wrong]
      - Visual analogue scale, 0-100:
        - [ability / inability <------------------------> (good luck/bad luck]
- Alternative knowledge probe (from Nagel et al., 2013)
  - **"In your view, which of the following sentences better describes Emma's situation?"**
    - Binary: ["Emma knows that she chose a necklace made of diamonds." OR "Emma feels like she knows that she chose a necklace made of diamonds, but he doesn't actually know that it is."]

Squirrel/Darrel Vignette (Turri et al., 2015)

For the Darrel vignette, all participants first read:

> 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 Zone 3. While scanning the river valley with his binoculars, Darrel sees a small, 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. Darrel records in his journal, ‘At least one red speckled ground squirrel in Zone 3 today.

- In the **knowledge** condition, participants then read:

    Ecologists are unaware that a complex network of aquifers recently began drying up in the park. These aquifers carry vital nutrients to the trees and other forms of plant life that support the squirrels. And the aquifers in the river valley running through Zone 3 are no exception. The animal Darrel is looking at is indeed a thirsty red speckled ground squirrel.

- In the **ignorance** condition, participants then read:

    Ecologists are unaware that a non-native species of prairie dog recently began invading the park. These prairie dogs also have red markings on their chest and belly. When these prairie dogs tried to invade Zone 3, the red speckled ground squirrels were unable to completely drive them away. And, the animal Darrel is looking at is indeed one of the prairie dogs.

- In the **Gettier** condition, participants then read:

    Ecologists are unaware that a non-native species of prairie dog recently began invading the park. These prairie dogs also have red markings on their chest and belly. When these prairie dogs tried to invade Zone 3, the red speckled ground squirrels were unable to completely drive them away. Still, the animal Darrel is looking at is a red speckled ground squirrel.

Measured Variables for Squirrel/Darrel

Presented in this order:

- Primary knowledge probe (from Turri et al., 2015):
    - **“Darrel ______ that there is at least one red speckled ground squirrel in Zone 3 today.”**
        - Visual analogue scale, 0-100:
            - [only believes <------------------------> knows]
- Comprehension question (from Turri et al., 2015):
    - **“Darrel is looking at a _____.”**
        - Binary: [ground squirrel/prairie dog]
- Reasonableness probe (from Turri et al., 2015):
    - **“It is _____ for Darrel to think that he is looking at a red speckled ground squirrel.”**
        - Visual analogue scale, 0-100:
            - [unreasonable <------------------------> reasonable]
- Luck/Ability probe (from Turri, 2016b)
    - **“Darrel got the ______ answer because of his _____.”**
        - Requires two responses:

- Binary: [right/wrong]
- Visual analogue scale, 0-100:
  - [ability/inability <-------------------------> (good luck/bad luck]

- Alternative knowledge probe (from Nagel et al., 2013)
  - **“In your view, which of the following sentences better describes Darrel’s situation?”**
    - Binary: ["Darrel knows that the animal he saw is a red speckled ground squirrel." OR "Darrel feels like he knows that the animal he saw is a red speckled ground squirrel, but he doesn't actually know that it is."]

Appendix C

Demographic, Exclusion, Covariate, and Study Experience Questions

Demographics/Covariates

- Age (open-ended response)
    - "How old are you? (in years)"
- Gender (drop-down box)
    - "What is your gender?"
        - Male, female, or other (open-ended)
- Race/Ethnicity (drop-down box)
    - "What is your ethnicity/race?"
        - *Version A (Main Protocol Version)*:
            - Used by all sites except where listed below.
                - White / European descent, Black / African descent, Latino*a / Latin American descent, Australian descent, Asian Southeast Asian descent, Native American, Hawaiian descent / Pacific Islands, Other (open-ended)
        - *Version B (Qualtrics Version):*
            - When the Qualtrics version was made, the response options were altered by that student team to reflect their sample. Teams that then used the Qualtrics version copied those response options.
            - Used by the following data collection sites in Qualtrics: AC2055, AC2054, and ACTURK
                - White/European, Black/African American, Hispanic Latino, East or Southeast Asian / Pacific Islander (e.g., from Japan, China, Korea, Vietnam, Thailand, Philippines, native Hawaiian), South Asian (e.g., from India, Pakistan), I prefer not to answer this question, Other (open-ended)
        - *Version C (Australian Version)*
            - One Australian team (AC206) altered the ethnicity options to better reflect Australian ethnic categories:
                - European descent, African descent, Latino*a / Latin American descent, Indigenous Australian or Torres Strait Islander descent, East Asian descent, South Asian descent, Pacific Island descent, Native American descent
        - *Version D (Russian Version)*
            - The Russian team (AC2053) altered the ethnicity options to better reflect Russian ethnic categories:
                - Russians, Ukrainians, Belarusians, Tatars, Armenians, Georgians, Kazakhs, Jews, Kyrgyz, Uzbeks, Tajiks, Chuvash, Other (fill in)
        - *Version E (Open-ended Version)*
            - Three student teams (at the same collection site) opted to only provide an open space for participants to fill in their ethnicity: AC1921, AC1921_NS, AC1921_S

- *Version F (Did Not Ask)*
    - Two sites opted to not ask participants to report their ethnicity at all: POL_001, AC2066
- Geopolitical region of residence (open-ended)
    - "What country do you currently live in?"
- Geopolitical region of birth (open-ended)
    - "What is your country of birth?"
- Education (drop-down box)
    - For all labs except two, this question was presented as follows:
        - "How many years did you attend school?"
            - Drop-down box with options 1 through 17.
    - For two labs (AC2054 and ACTURK), this question was presented as follows:
        - "What is your highest level of education attained?"
            - Drop-down box with these options: Less than high school, high school diploma (or GED), some college or a 2-year college degree (A.A.), 4-year college degree (B.A., B.S.), Master's degree (M.A., M.S.), Graduate or professional degree (J.D., Ph.D., M.D.)

Exclusion Questions

- Self-assessed language proficiency
    - "How well do you speak [*insert survey language here*]?"
        - Very well, well, not very good, not good at all
- Study purpose
    - "What do you think is the purpose of this study?"
- Impression of materials
    - "What was your impression of the materials in this study?"
- Previous participation in a similar study
    - "Have you ever participated in a similar study?"

The Study Experience Questionnaire

The following questionnaire is your chance to give feedback on the study you have just participated in.

Please use the following anchors to describe your experience of this study.
Please circle the number that best represents your experience of the study relative to the two ends of the scale. Note that a '5' is the middle of a scale and can be used if you are not sure of an answer.

How much did you enjoy the study?

I enjoyed the study a lot Not sure I did **not** enjoy the study at all

1 2 3 4 5 6 7 8 9

How nervous were you during the study?

I was very nervous during the study | Not sure | I was **not** nervous during the study at all

1 2 3 4 5 6 7 8 9

How difficult did you find the study?

I found the study tasks very difficult to complete | Not sure | I did **not** find the study tasks difficult to complete at all

1 2 3 4 5 6 7 8 9

How boring did you find the study?

I found the study task very boring | Not sure | I did **not** find the study activity boring at all

1 2 3 4 5 6 7 8 9

How tiring did you find the study?

I found the study task very tiring | Not sure | I did **not** find the study task tiring at all

1 2 3 4 5 6 7 8 9

How quickly did you adjust to the study task?

I was able to adjust to the study task very quickly

Not sure

I was **not** able to adjust to the study task quickly

1 2 3 4 5 6 7 8 9

How regularly do you take part in research studies?

I have taken part in many research studies

Not sure

I have **never** taken part in a research study before

1 2 3 4 5 6 7 8 9

How self-conscious of your responses were you during the study?

I was very self-conscious of the responses I gave in this study

Not sure

I was **not at all** self-conscious of the responses I gave in this study

1 2 3 4 5 6 7 8 9

How motivated were you to help the researchers during the study?

I was strongly motivated to help make the study a success for the researchers

Not sure

I was **not** at all motivated to help make the study a success for the researchers

1 2 3 4 5 6 7 8 9

To what extent did you believe you were contributing to important research?

I believe that my participation was contributing to very important research

Not sure

I **do not** believe that my participation was contributing to important research

1 2 3 4 5 6 7 8 9

To what extent were you trying to work out the aim of the study during your participation?

I was trying to work out the aim of the study during my participation

Not sure

I was **not** trying to work out the aim of the study during my participation

1 2 3 4 5 6 7 8 9

Do you have any further comments about your experience of this study that we have not addressed above? Please give any further comments about this study below:

Appendix D

Exclusion Ratings

How exclusions were made

- No or NA gets 0 points; Maybe gets 1 point; Yes/test gets 2 points; Participants are marked as "excluded" if they get 4 total points across three coders.
- Participants are also marked as "nonsense" if they did not write a legible answer or simply typed gibberish. These data points are not excluded but marked.

Instructions Given to Raters

- *Note: These instructions were adapted from instructions written by William McAuliffe and Hannah Moshontz for an unrelated project*

  We need help coding open-ended responses that will inform our pre-registered exclusion criteria for this project. Specifically, we will exclude data on the basis of a suspicion check (whether people guess the study hypothesis) and previous study participation (whether people describe having participated in similar studies before).

  All participants were asked two questions (with some labs asking slight variations): What is, in your opinion, the purpose of this study? (purpose) Have you ever participated in a similar study? If yes, please describe the study. (previous)

  Your task is to evaluate people's answers to these questions. We will have 2 people evaluate every response and then we will exclude people based on the average.

  For each question this is how we would like you to evaluate answers. If you are coding from a language other than English, please directly assess the question (rather than translating it) and provide a code/label in English (yes, maybe, no, test, as described below).

- **Purpose**
  - *Yes*
    - The participant identified that we are studying justified true belief and gettier cases.
      - Example "yes" coding cases: "To test exceptions to the Justified True Belief theory"
  - *Maybe*
    - The participant describes something similar to the true study hypothesis (true knowledge is different from a lucky or incidentally correct belief).
      - Example "maybe" coding cases: "To see if a story can change ones perception of knowledge based on luck or ability"
  - *No*
    - The participant did not identify the study hypothesis or offer a very vague description, which might include the words belief or knowledge.
      - Example "no" coding cases: "I think the purpose was to see how do people classify if someone knows something or if they just

strongly agree with it"; OR "understand how people view scenarios based on the words used to describe them"

- *Test*
    - The response indicates that it is a test case
        - Example "test" coding case: "TEST"; OR "test"; OR "this is a test"
- *NA*
    - If you are unsure how to code a response, you can write NA.
        - Example "NA" coding case: "nnnnnnnnnn"

- **Previous**
    - *Yes*
        - The participant has participated in this exact study before, or an exact replication of it.
            - Example "yes" coding cases: "Yes, I completed this study before."
    - *Maybe*
        - The participant has participated in a similar study before, or may have based on their description.
            - Example "maybe" coding cases: "Yes another study that was very similar."; "Yes, I have participated in a study that asked similar questions but had slightly different scenarios"
    - *No*
        - The participant has not participated in this study or a similar study before based on their description.
            - Example "no" coding cases: "nope"; "Yes, I have participated in a study for course credit before."'; "Yes, I have done studies where I read scenarios and answered questions about them."
    - *Test*
        - The response indicates that it is a test case
            - Example "test" coding case: "TEST"; OR "test"; OR "this is a test"
    - *NA*
        - If you are unsure how to code a response, you can write NA.
            - Example "NA" coding case: "nnnnnnnnnn"

- **Do's and Don'ts** Take breaks! This work is hopefully interesting, but it can be cognitively exhausting. If you are having trouble paying attention while you are doing this or if you feel tired of it, please take a break.

- After you label a response, do not go back and change it later. This may be tempting to do after mentally comparing how you rated different responses, but just carefully work through each response and know that your initial rating is final.

- Don't discuss your ratings with other raters—this will invalidate everyone's work. Do assign labels for every response in your assigned sheet(s). If you would like to contribute more, please email the person listed at the top of this sheet.

- **To summarize, for each set of answers** Read the answer to the question and assign a label that describes either whether people intuited the study hypothesis (for purpose) or whether people participated in a similar study previously (for previous)
- **Coding form includes** [id] A subject id number [survey_lang] [purpose/previous] The answer people gave to the question [code] The code/label that you are assigning to the answer (yes, no, maybe, test)

Appendix E

Extension Vignettes

Art/Julie Vignette

An expertise extension was added after the primary protocol of the study by three collection sites (representing 10 samples: AC1804, AC1805, AC1805_N, AC1807, AC1810, AC1808, AC1809, AC1907, AC1907, and AC1940). In this extension, protagonist expertise (two levels: expert and novice) was manipulated in addition to the primary manipulation of propositional knowledge (three levels: knowledge control, gettier case, and ignorance control). Specifically, after participants completed the main portion of the approved protocol, they were then randomly assigned to receive one of six versions (i.e., 3 propositional knowledge conditions x 2 expertise conditions) of a new vignette created specifically for the proposed extension (Art/Juli). After reading their one randomly assigned vignette, participants then completed dependent measures adapted from the main protocol of the study.

Expert Conditions

For the expert Julie vignette, all participants first read:

> Julie is an experienced art appraiser who is visiting a museum with her friend to see a special exhibit of rare works of art. Upon entering the museum, a particular painting catches her eye. Julie analyzes the painting's overall appearance, brush strokes and technique. She examines the signature at the bottom of the painting and tells her friend, "This painting was done by Monet."

- In the **knowledge** condition, participants then read:

  > What Julie does not realize is that a collection of Monet's paintings was recently found in a residential estate. These paintings were kept in pristine condition and include some never before seen works done by Monet. Julie has never encountered one of these newly discovered works as they have not been available to the public. When she tells her friend, "This painting was done by Monet," the painting she is looking at is indeed a real work of art done by Monet.

- In the **ignorance** condition, participants then read:

> What Julie does not realize is that there is a forger in the area who has been illegally duplicating many esteemed artists' works, including some done by Monet. This forger is very talented and his works have been making their way into local exhibitions. And, when Julie tells her friend, "This painting was done by Monet," the painting she is looking at is indeed a fake work of art done by the forger, not by Monet.

- In the **Gettier** condition, participants then read:

> What Julie does not realize is that there is a forger in the area who has been illegally duplicating many esteemed artists' works, including some done by Monet. This forger is very talented and his works are starting to make their way into local exhibitions. Still, when Julie tells her friend, "This painting was done by Monet," the painting she is looking at is indeed a real work of art done by Monet.

Novice Conditions

For the novice Julie vignette, all participants first read:

> Julie is visiting a museum with her friend to see a special exhibit of rare works of art. Upon entering the museum, a particular painting catches her eye. Julie analyzes the painting's overall appearance, brush strokes and technique. She examines the signature at the bottom of the painting and tells her friend, "This painting was done by Monet."

- In the **knowledge** condition, participants then read:

> What Julie does not realize is that a collection of Monet's paintings was recently found in a residential estate. These paintings were kept in pristine condition and include some never before seen works done by Monet.
>
> Julie has never encountered one of these newly discovered works as they have not been available to the public. When she tells her friend, "This painting was done by Monet," the painting she is looking at is indeed a real work of art done by Monet.

- In the **ignorance** condition, participants then read:

> What Julie does not realize is that there is a forger in the area who has been illegally duplicating many esteemed artists' works, including some done by Monet. This forger is very talented and his works have been making their way into local exhibitions. And, when Julie tells her friend, "This painting was done by Monet," the painting she is looking at is indeed a fake work of art done by the forger, not by Monet.

- In the **Gettier** condition, participants then read:

> What Julie does not realize is that there is a forger in the area who has been illegally duplicating many esteemed artists' works, including some done by Monet. This forger is very talented and his works are starting to make their way into local exhibitions. Still, when Julie tells her friend, "This painting was done by Monet," the painting she is looking at is indeed a real work of art done by Monet.

Measured Variables for Art/Julie (Presented in this Order)

- Primary knowledge probe (based on Turri et al., 2015):
    - **“Julie ______ that she is looking at a real painting done by Monet.”**
        - Visual analogue scale, 0-100:
            - [only believes <------------------------> knows]
- Comprehension question (based on Turri et al., 2015):
    - **“Julie is pointing at a _____. ”**
        - Binary: [real painting done by Monet/fake painting done by a forger]
- Reasonableness probe (based on Turri et al., 2015):
    - **“It is _____ for Julie to think that she is pointing at a real painting done by Monet.”**
        - Visual analogue scale, 0-100:
            - [unreasonable <------------------------> reasonable]
- Luck/Ability probe (based on Turri, 2016b)
    - **“Julie got the ______ answer because of her _____.”**
        - Requires two responses:
            - Binary: [right/wrong]
            - Visual analogue scale, 0-100:
                - [(in)ability<------------------------> (good/bad) luck]
- Alternative knowledge probe (based on Nagel et al., 2013)
    - **“In your view, which of the following sentences better describes Julie’s situation?”**
        - Binary: ["Julie knows that the painting she is pointing at is a real painting done by Monet." OR "Julie feels like she knows that the painting she is pointing at is a real painting done by Monet, but she doesn't actually know that it is."]