Registered Replication Report: Turri, Buckwalter, & Blouw (2015)

Braeden F. Hall, Jordan R. Wagge, Gerit Pfuhl, Stefan Stieger, Evie Vergauwe, Hans IJzerman, Waldir Monteiro Sampaio, Manyu Li, Martin Voracek, Joaquín Morís, Andree Hartanto, Sophia Weissgerber, Christoph Schild, Gilad Feldman, Leslie D. Cramblet Alvarez, Elena Kelsey Henderson, Florian Cova, Amy T. Nusbaum, Ulrich Tran, Martin J. Mækelæ, Sau-Chin Chen, Sümeyra Özben, Krystian Barzykowski, Peter Szecsi, Gwenael Kaminski, Nikki Legate, Anthony J. Krafnick, Erica D. Musser, Felizitas Pernerstorfer, Colleen Lamb, Balazs Aczel, Asil A. Özdoğru, Carlota Batres, Gerrit Heßmann, Kaitlyn M. Werner, Ewout H. Meijer, Daniel Storage, Michael Schulte-Mecklenbeck, Nicole Mathis, Yanna Weisberg, Vilius Dranseika, Erin M. Buchanan, Michael Andreychik, Jack D. Arnal, Gerald J. Haeffel, Thomas Rhys Evans, Sean T. H. Lee Teck Hao, Andy P. Field, Anastasiia Tkalich, David Redman, Marianne Fallon, Sara Alvarez Solas, Victor Webersberger, Lee Ru Xuan, Julia Teeter, Mirian Fergnani, Chiu C. Peinn, Teresa Nitsche, Sydney S. Thelen, Yannick Michael Endres, Neil W. Kirk, Carmel A. Levitan, Justas Žilys, Evin Yildirim, Sylwia Adamus, Tina Grünemay, Cara Sibert, Molly Metz, Luisa Joel, Claudia Ferreira, Jana Mertsch, Celine Rognskaug, Yu-Hsuan Ku, Maria C. Vetter, Lim Ke Ying, Corey L. Fincher, Natasha Godkin, Kira Aschenbroich, Lea Müller, Elena Alexandra Magazin, Eleanor H. Lyman, Catie Norwood, Kuba Dubis, Erin M. Isola, Samantha Schwegmann, Paula Bange, Ian Buhmann, Kaitlyn Larkin, Ngoc Phuong Trinh Nguyen, Paweł Lubomski, Caroline Kolle, Robert Calin-Jageman, Gianni Ribeiro, Shumin Liu, Karolina Golik, Anna Kulpe, Janey M. Almaraz, Caitlin R. Eckerman, Chia-Shien Lin, Sarah Crain, Nicholas J. DeLollo, Russell Warne, Hannah Johnson, Kyla N. Lewis, Brady Wiggins, Radosław Kabut, Marton Kovacs, Mary K. Walsh, Selina Bornmann, Julita Kielińska, Alexis Martel Lamothe, Annie Hanh Vu, Mark J. Brandt, Christopher R. Chartier, & Jon E. Grahe

Accelerated CREP

Multilab close replication of: Experiment 1 from Turri, J., Buckwalter, W., & Blouw, P. (2015). Knowledge and luck. *Psychonomic Bulletin and Review*, *22*, 378-390.

Data and registered protocols: https://osf.io/n5b3w/

Abstract

Justified True Belief (JTB) theory defines a person's ability to know something as having a belief that is both justified and true (i.e., knowledge is justified true belief). However, in 1963, Gettier argued that JTB is insufficient because it does not account for scenarios, called Gettier cases, wherein a person is justified for believing something true but only because of luck. Lay people's intuitions about knowledge may lead them to believe individuals in these cases lack knowledge (referred to as Gettier intuitions), making luck an unreasonable justification for belief. We aim to provide a robust estimate of the Gettier intuition effect size by replicating Turri and colleagues' (2015) Experiment 1. The Collaborative Replications and Education Project (CREP) selected this study for replication based on its undergraduate appeal, feasibility, and pedagogical value. Considering some inconsistent results, suboptimal designs, and varying evidence for cultural variation (e.g., Machery et al., 2015; Seyedsayamdost et al., 2015; Weinberg et al., 2001), the superior methodology of Turri et al. (2015) also makes it an important study to replicate cross-culturally. Therefore, we propose a Registered Replication Report of Turri and colleagues' (2015) Experiment 1 (49 labs from 22 countries across 5 continents signed up at time of submission; expected minimum N = 2,000). Results of this study are expected to provide a clearer picture of Gettier intuitions and lay people's theory and practice of knowledge. The data are released in two phases according to a predefined plan to facilitate exploratory cross-cultural analyses. Preprint: psyarxiv.com/zeux9 Preregistered protocols: osf.io/n5b3w/

Keywords: Folk epistemic intuitions, Beliefs, Social cognition, Gettier intuitions, Justified True Belief, Multilevel modeling, Multilab, Replication

Registered Replication Report: Turri, Buckwalter, & Blouw (2015)

Justified True Belief and the Gettier Problem

By some accounts, the Justified True Belief (JTB) theory of knowledge (or alternative versions of it) has been an important account of propositional knowledge in Western thinking for the past two millennia (e.g., Jacquette, 1996; Moser, 2002; but cf., Dutant, 2015; Turri, 2016a). The JTB account states that a person's claim can only be 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, but also have sufficient reason for believing the claim to be true. Thus, making a lucky guess that happens to reflect the truth is not enough to say that you know something. This raises questions regarding the evidence of universal *epistemic intuition* (i.e., intuitions about knowledge) and its underpinnings.

In 1963, Gettier challenged the sufficiency of JTB to account for all knowledge by presenting two strong counterexamples that are inconsistent with predictions derived from JTB. These counterexamples (i.e., *Gettier 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. For example:

> Two men, Smith and Jones, have applied to the same job at the same company. Much to Smith's disappointment, the president of the company has told Smith that Jones will ultimately get the job. Smith then notices that Jones has ten coins in his pocket, coins

> which Smith counted himself (oddly enough). Smith then infers that the man who gets the job (who he assumes will be Jones) will have ten coins in his pocket (which he counted in Jones' pocket himself): a belief that is well founded by the evidence and therefore justified. However, quite unexpectedly, Smith ends up getting the job! And, unbeknownst to himself, Smith coincidentally has ten coins in his pocket too. Although this was not the outcome that Smith was expecting, his inferred belief that the man who has ten coins in his pocket will get the job still turned out to be true, just not for the reason he thought (Gettier, 1963).

Even though Smith's belief was both true and justified, Gettier argued that Smith does not have knowledge in this case - Smith just got lucky. Instances in which people's epistemic intuitions do not lead them to attribute knowledge to a Gettier protagonist have since been referred to as *Gettier intuitions* (DePaul & Ramsey, 1998; Machery et al., 2017; Sosa, 2007). However, the extent to which people demonstrate Gettier intuitions remains unclear.

Gettier Intuitions: Important, but Inconclusive Evidence

Results from an experimental philosophy study by Turri et al. (2015; Experiment 1) suggest that lay people do not demonstrate Gettier intuitions. In this study, participants were asked whether a protagonist in one of three stories knew or only believed a claim. In the experimental Gettier case 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 amidst many animals of a different, almost identical species (i.e., counterfeit species). The other two stories presented the same scenario with slight changes: in the knowledge control, the story never mentions the other identical species (i.e., no counterfeit) and in the ignorance control, the protagonist incorrectly identifies the counterfeit species as the target species. Turri et al. (2015) compared the rate at which participants attributed knowledge to the protagonist in the Gettier case to the rates participants attributed knowledge in the control stories.

Results demonstrated that participants did not attribute knowledge to the protagonist in the Gettier case at rates significantly different than the knowledge control - which suggests that the luck involved in a faulty justification that still leads to a true claim is consistent with lay people's conception of knowledge (Turri et al., 2015). In contrast, other research has found that lay people are in fact sensitive to such luck; thereby, demonstrating Gettier intuitions (Nagel, San Juan, & Mar, 2013). These inconsistencies highlight the gaps in what we currently know about the conditions necessary for people to attribute knowledge to others.

The precise magnitude of the Gettier intuition effect is currently unknown. Experimental studies have demonstrated that knowledge attribution rates for different Gettier cases vary from lower than 20% (Gettier intuition supported) to higher than 80% (Gettier intuition not supported; Turri, 2016a). Such inconsistencies are perhaps due to two major reasons: (1) people rely on different epistemic intuitions to make judgements about the various types of Gettier cases studied in the literature (e.g., "counterfeit object" cases, "authentic-evidence" cases, "apparent-evidence" cases, etc.) and (2) suboptimal experimental designs, including lack of matched controls and underpowered samples (see Colaço et al., 2014; Machery et al., 2017; Nagel, Mar, & San Juan, 2013; Nagel, San Juan, & Mar, 2013; Powell, Horne, & Pinillos, 2013; Powell, Horne, Pinillos, & Holyoak, 2015; Starmans & Friedman, 2012; 2013; Turri et al., 2015; Weinberg et al., 2001). Our project focuses on obtaining a more precise estimate of the effect size of Gettier intuitions as they relate specifically to "counterfeit object" type Gettier cases. Therefore, we propose a large, high-powered, and cross-country replication of Turri et al. (2015)'s Experiment 1 (a "counterfeit object" Gettier case). We believe that this study serves as a particularly good paradigm for studying Gettier intuitions, making it a suitable candidate for such a large-scale replication.

Although the overall literature on epistemic intuitions has demonstrated inconsistencies across different types of Gettier cases, there is no evidence that this is true for the particular type of Gettier case ("counterfeit object" type Gettier case) studied in Turri et al. (2015; Experiment 1). Previous findings by Nagel, San Juan, & Mar, 2013 may contradict Turri and colleagues' findings, but results from this study have been called into question. In a reply to this study, Starmans and Friedman (2013) pointed out that: (1) Nagel, San Juan, & Mar (2013) employed a questioning method that biased participants to deny knowledge, (2) careful examination of participants' responses revealed that they did in fact attribute knowledge to protagonists in Gettier cases, and (3) Nagel, San Juan, & Mar (2013) misconstrued the distinction between 'apparent' and 'authentic' evidence, and used scenarios that did not feature the structure that characterizes most Gettier cases (but cf. Nagel, Mar, & San Juan, 2013). Starmans and Friedman (2013) argued that Nagel, San Juan, & Mar's (2013) findings are fully compatible with the claim that lay people do generally attribute knowledge in Gettier cases (Gettier intuition not demonstrated; but cf. Nagel, Mar, San Juan, 2013) - which we plan to test in the current study.

The Current Study

We are planning a large Registered Replication Report (RRR) following the design of Experiment 1 in Turri et al. (2015). As in Turri et al. (2015), the current study seeks to test the effects of a protagonist making an inference from a false belief that is true by unknowingly and luckily choosing a true, genuine object among many convincing counterfeits on participants' knowledge attribution: people will judge a person's belief as knowledge when the belief is justified and true ("No Threat"; or JTB case) not significantly differently than when the belief is justified and only luckily true (“Threat”; or Gettier case), and statistically less often when the justified belief is false (“No Detection”; or Ignorance case; see Figure 1 for original results).

Second, we will test whether people say the protagonist in each condition is reasonable for believing what they believe. Following Turri et al. (2015), we will examine whether differences in knowledge attribution rates are due to perceived differences of what is reasonable for the protagonist to believe (i.e., is the given protagonist justified in their belief?; see Figure 1 for original results). We will also attempt to replicate Turri and colleagues’ (2015) test of whether the number of people who attribute knowledge to the given protagonist in each of these cases differs from the number of people we would expect to attribute knowledge in each case based on chance alone.

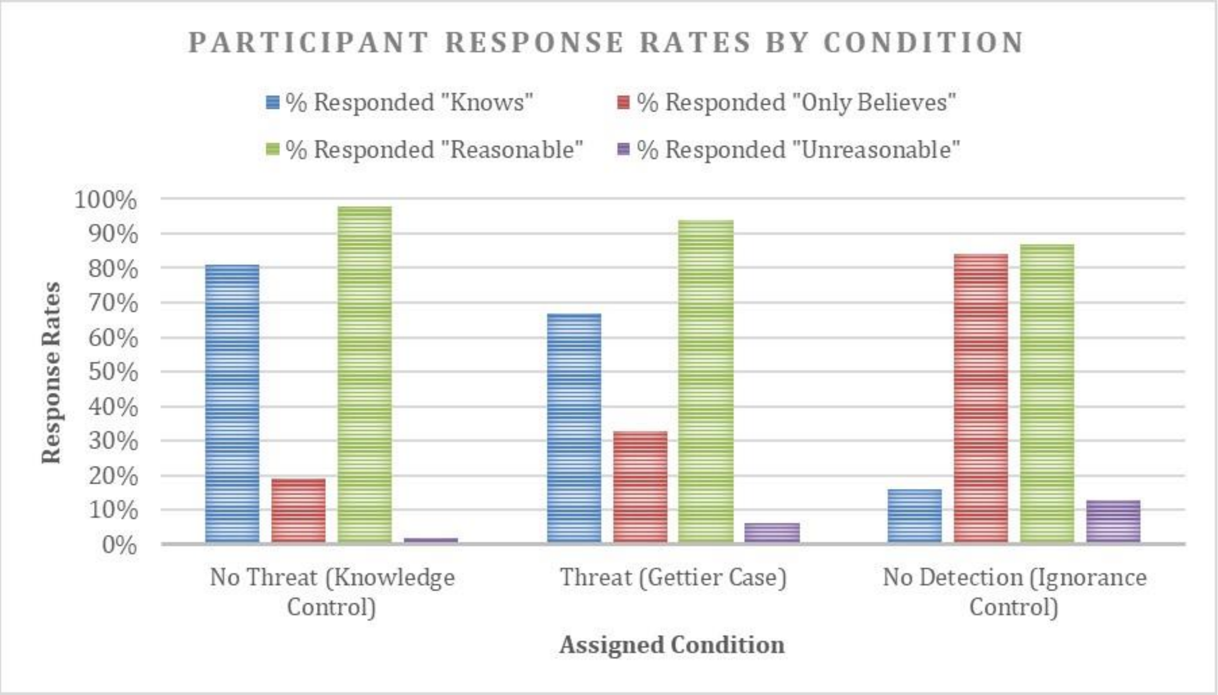


Figure 1. Bar graphs depicting the original response rates from Turri, Buckwalter, and Blouw’s (2015) findings in Experiment 1. Participants in each condition were asked whether Darrel

“knows or only believes” that he saw a red-speckled ground squirrel and whether it was “reasonable or unreasonable” for Darrel to say that he saw a red-speckled ground squirrel.

To improve our test of the extent to which Turri et al.’s (2015) findings replicate across different procedures, scenarios, and cultures, we present: (1) several methodological considerations (related to design, measurement, and culture) raised by past experimental philosophy research that may account for why prior Gettier intuition findings have demonstrated variation, and 2) the corresponding modifications we will make to the original Turri et al. (2015) study design to address these concerns.

Design considerations.

The current consensus for these inconsistent findings is that the established vignettes may elicit different intuitions from the general public based on the particular structure of the tested Gettier case (Turri, 2016a). The two original counterexamples Gettier used in his 1963 paper both feature an “inference-from-false-belief” type structure wherein a protagonist forms an initial justified but false belief from which a true claim is then inferred, much like the following scenario typically described in the literature (e.g., Nagel et al., 2013): Person *S* reasonably (yet incorrectly) believes that, "This shop sells only real diamonds". Person *S* then makes an inference from this justified false belief that because "this diamond is in this shop", then “this diamond is a real diamond”. However, the shop in question actually mostly only sells cubic zirconium stones, dishonestly designed to look just like diamonds. Therefore, Person *S*’s original justified belief that “This is a shop that sells only real diamonds” is false. Despite this false premise though, Person *S* unwittingly chooses to buy the one and only real diamond in the store.

So, while the initial justified belief ("This shop sells only real diamonds") turned out to be false, the resulting inferred belief ("This diamond is a real diamond") still turned out to be a true claim.

Although most philosophers now typically define Gettier cases as any instance that is intended to illustrate the non-equivalence of justified true belief and knowledge (where the case is viewed as not being consistent with knowledge), others have used it 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 (reviewed in Nagel, San Juan, & Mar, 2013). Instead of defining Gettier cases as instances that are intended to show a disparity between justified true belief and knowledge, as Nagel and colleagues (2013) do, we instead operationalize Gettier cases as scenarios of the inference-from-false-belief type structure featured in Gettier's original article.

Due to the problems that arise from ignoring the kind of stimulus variation present in the experimental philosophy literature (Clark, 1973; Judd & Westfall, 2012; Kenny, 1985; Nagel, San Juan, & Mar, 2013; Starmans & Friedman, 2013; Wells & Windschitl, 1999; Westfall, Judd, & Kenny, 2015), we will also attempt to conceptually replicate the original Turri et al. (2015) finding using additional "counterfeit object" type Gettier vignettes from the literature, wherein a protagonist makes an inference from a false belief that is true by unknowingly and luckily choosing a true, genuine object among many convincing counterfeits (e.g., the fake barn or "Emma" vignette adapted from Colaço et al., 2014; the counterfeit diamond or "Gerald" vignette adapted from Nagel, San Juan, & Mar, 2013; see stimulus sampling plan below for more details). Doing so will allow us to test the generalizability of Turri et al.'s (2015) Experiment 1 manipulation to other similar "counterfeit object" cases (and to reduce sampling error). This will be our main test.

Measurement considerations.

Recent research (Turri, 2016c) relying on a more sensitive scaled measure of knowledge attribution found a statistically significant difference between an appropriately matched knowledge control condition and a *Gettier* experimental condition. In our correspondence (Hall et al., 2018), Turri stated that the observed non-significant difference between the knowledge control and the Gettier case condition (X^2 (2, N = 98) = 2.63, p = .164, Cramér's V = .164; Gettier intuition not supported) may not have been significant due to the binary format of the knowledge probe used and the study's underpowered sample size. Turri also noted that closely matched knowledge control and Gettier condition comparisons do sometimes reveal a small statistically significant difference in the overall literature (Hall et al., 2018).

To address these concerns, we will use a visual analogue scale (0-100) in lieu of the original binary (i.e., knows/only believes) response variable, because the use of visual analogue scales (VAS) may be as efficacious as Likert-type response scales, while they provide somewhat more fine-grained data for analysis via parametric statistics (Bishop & Herron 2015). Although using a VAS departs from the original study, and also, from how people typically make these kinds of judgments in ordinary life, our pretest using a VAS format suggests that people did not have difficulty intuitively answering the questions correctly (knowledge controls and ignorance controls demonstrated paradigmatic rates, see Appendix B).

Further differences in knowledge attribution have also been found based on how subjects are asked whether a target has knowledge (e.g., "does the target know?" *vs*. "does the target know, or do they only think they know?"; e.g., Machery et al., 2017; Nagel et al., 2013). To check for these differences in knowledge attribution based on the form of the knowledge question, we will ask an exploratory binary knowledge attribution question *after* the visual analogue scales (see Appendix B). After the alternative knowledge probe, we will also ask an additional exploratory question related to the perception of luck and ability that may moderate the effect (e.g., "Darrel got the [wrong/right] answer because of his [inability/ability/good luck/bad luck]"; Nagel et al., 2013; Turri, 2016a; see Appendix B).

Cultural considerations.

The literature has demonstrated some cultural variations in knowledge attribution (e.g., Buckwalter & Stich, 2010; Kim & Yuan, 2015; Machery et al., 2015; 2017; Nagel, San Juan, & Mar, 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 tended to 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 similarly matched controls) found no such cultural differences (e.g., Machery et al., 2015, 2017; Seyedsayamdost, 2015). In one of the largest of these cross-cultural studies, Machery and colleagues (2015) provided evidence that people exhibit Gettier intuitions across quite different cultures and languages (i.e., USA, Brazil, India, and Japan), suggesting a "species-typical core folk epistemology" wherein justification, truth, and belief are insufficient for people to attribute another person with the concept they express by the words that translate "to know" (Machery et al., 2015, pp. 12).

Past findings showcase the importance of utilizing control conditions and closely matched stimuli (Turri, 2016a). While more recent studies have utilized 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 & Yaun, 2015; Machery et al., 2015; 2017; Seyedsayamdost, 2015). For example, Machery and colleagues (2015) compared responses to control conditions but used entirely different vignettes with different protagonists for each condition. By contrast, Turri et al. (2015) used slight variations of the same vignette for each condition: “No Threat” (i.e., knowledge control), “Threat” (i.e., Gettier case), and “No Detection” (i.e., ignorance control). Because the vignettes used in Turri et al. (2015; see Appendix B) differ only in the words necessary to alter the condition of the protagonist’s belief (unlike some work, e.g., Buckwalter & Stich, 2010; Machery et al., 2015; Weinberg et al., 2001), this design is better suited than others for making inferences about why participants attribute knowledge to protagonists. Therefore, we also ensure that the two added vignettes (the “Fake Barn/Gerald” vignette and the “Diamond/Emma” vignette) are also tested alongside minimally matched pairs similar to those used in Turri et al.’s (2015) Experiment 1 (see Appendix B for full details).

In light of these considerations, we will test the Gettier intuition effect in a variety of countries (49 labs from 22 countries[1] across 5 continents[2]) while attempting to address methodological concerns (i.e., measurement sensitivity, lack of matched controls, stimulus variation). Such a multisite, cross-country replication could also provide a well-powered exploration of cross-cultural similarities and differences in how people attribute knowledge while utilizing control conditions and closely matched stimuli for several vignettes. We will therefore include a plan for a phased release of the data (⅔ upon acceptance of the paper, ⅓ six months after acceptance) to allow other researchers to do cross-cultural comparisons

[1] Australia, Austria, Brazil, Canada, Denmark, Ecuador, France, Germany, Hong Kong, Hungary, Lithuania, Norway, Poland, Republic of China, Singapore, Spain, Switzerland, Taiwan, The Netherlands, Turkey, United Kingdom, and United States of America.

[2] Australia, Asia, Europe, North America, and South America

Disclosures

Data, materials, and online resources: All data will be posted publicly on our master OSF page (https://osf.io/n5b3w/), and each contributing site will post their data on an OSF page linked to our master OSF page:

Reporting: "We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study" (see Simmons, Nelson, & Simonsohn, 2011).

Ethical approval: All contributing labs are required to submit their local institutional ethics approval prior to data collection as part of their pre-registration and CREP review process and will be carried out in accordance with the provisions of the World Medical Association Declaration of Helsinki. All participating labs will post their ethics approval to their lab's OSF page for this study.

Protocol Development

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), 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 “a distributed network of laboratories designed to enable and support crowdsourced research projects” with a mission “to accelerate the accumulation of reliable and generalizable evidence in psychological science” (Moshontz et al., 2018, p. 2). 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 countries.

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). Six months after calls for contributors, 49 labs from 22 countries have signed up to contribute data as of this submission. For the purposes of both quality control and educational value, we require that all participating labs pre-register their own independent direct replication protocol with a video of their methodology posted on their Open Science Framework (OSF) page. For quality control, these pages must be approved by our executive team to ensure that each lab meets all standards and procedures set forth in this protocol prior to data collection.

Contributing laboratories are required to obtain any necessary ethics approval from their institutions, with approvals also posted to their OSF page. Once ethics, protocol, and taped methodology have been approved by the CREP team, contributors will be allowed to begin data collection starting March/April/May 1st, 2019 (depending on date of in principle acceptance). Contributors may continue to sign on to participate in this study until February 1st, 2019. We chose this date to ensure that teams will have at least three months to get ethics approval, get CREP pre-registration approval, and collect and submit their report. Overall data collection will end on April 1st, 2020. Data release is dependent upon manuscript acceptance, but full data release will be six months after the first ⅔ of the data release. The overall protocol for this direct replication of Experiment 1 of Turri et al. (2015) was developed by all co-authors with suggestions from one of the original authors (i.e., Turri; see Hall et al., 2018 for his comments).

Protocol Requirements

Sampling plan.

Student teams from any country are invited to collect samples. 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. However, online data collection services that recruit subjects who are then paid for participating, such as Amazon Mechanical Turk (MTurk), will not be permitted (see below for rationale behind this restriction). We will, however, independently collect one large (*n* = 500) MTurk sample of US participants to compare a sample comparable to Turri et al.'s (2015) with the rest of the studied samples (more about this below). Samples may not be drawn from institutions comprised of vulnerable populations (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.

We pretested two additional vignettes/stimuli (each with two controls) from the literature based on similarity to the replicated vignette (i.e., counterfeit object Gettier cases). Each site will be required to pre-register a stimulus sample size of 3 ("Darrel", "Gerald", and "Emma") and a participant target sample size of at least 40. Our goal is to collect data from at least 50 independent contributors. We are 98% towards 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 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 for errors.

Although we will sample from many different populations, results from recent multilab studies amongst students (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. Group vs. individual administration and compensated vs. not compensated will also be tested as test setting covariates. 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. The CREP is an educational project that 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 will pre-registered and collect 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 university samples, we will also encourage (but not require) sites to collect an additional non-university sample (*N* = 40) 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.

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. 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 Colaca 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 did not have matched knowledge and ignorance control conditions similar to Turri et al. (2015). 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. We will also record demographic information[3] 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, country of residence, country 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, we will instead ask

[3] 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.

how well participants speak the language in which they are being tested and if said language is their first language.

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. 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.

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. 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.

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 make any 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 and colleagues (2015) used in the original study (see Procedures). 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)[4]. 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.

[4] Resulting in 6 propositional knowledge condition order combinations, and randomizing order of presenting vignettes (2 possible order combinations), resulting in 12 possible flows in SoSciSurvey.

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 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 Turri et al. (2015) vignette should be presented first, and then the remaining two vignettes 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 (knowledge control, Gettier case, and ignorance control), giving rise to a cross-classified data structure. 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 comprehension question to control for understanding. 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 replicated questions will always be asked in the same order (knowledge/ comprehension/ reasonableness) for each vignette. However, unlike the original study, response options on the visual analogue scales (e.g., knows/only believes) will not be rotated randomly to control for order effects (e.g., sometimes "knows" is on left side of the scale, and sometimes "only believes" is on the left side). We are opting out of this procedure because neither the original study nor our pretests found an order effect (Hall et al., 2018; Turri et al., 2015).

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). 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. 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). 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 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).

Data collection stopping rules and exclusions.

Each site will pre-register a minimum target sample size of 40 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. 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 country 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.

Analysis Plan

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) with multilevel modeling (for a visualization of this data structure, see Figure 2). 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 first vignette they read will always be the original Turri et al. (2015) Experiment 1 vignette (i.e., the “Darrel” vignette) randomized into one of the three belief conditions. The remaining vignettes will then be presented in random order, each also randomized and counter-balanced to one of the three conditions. 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).

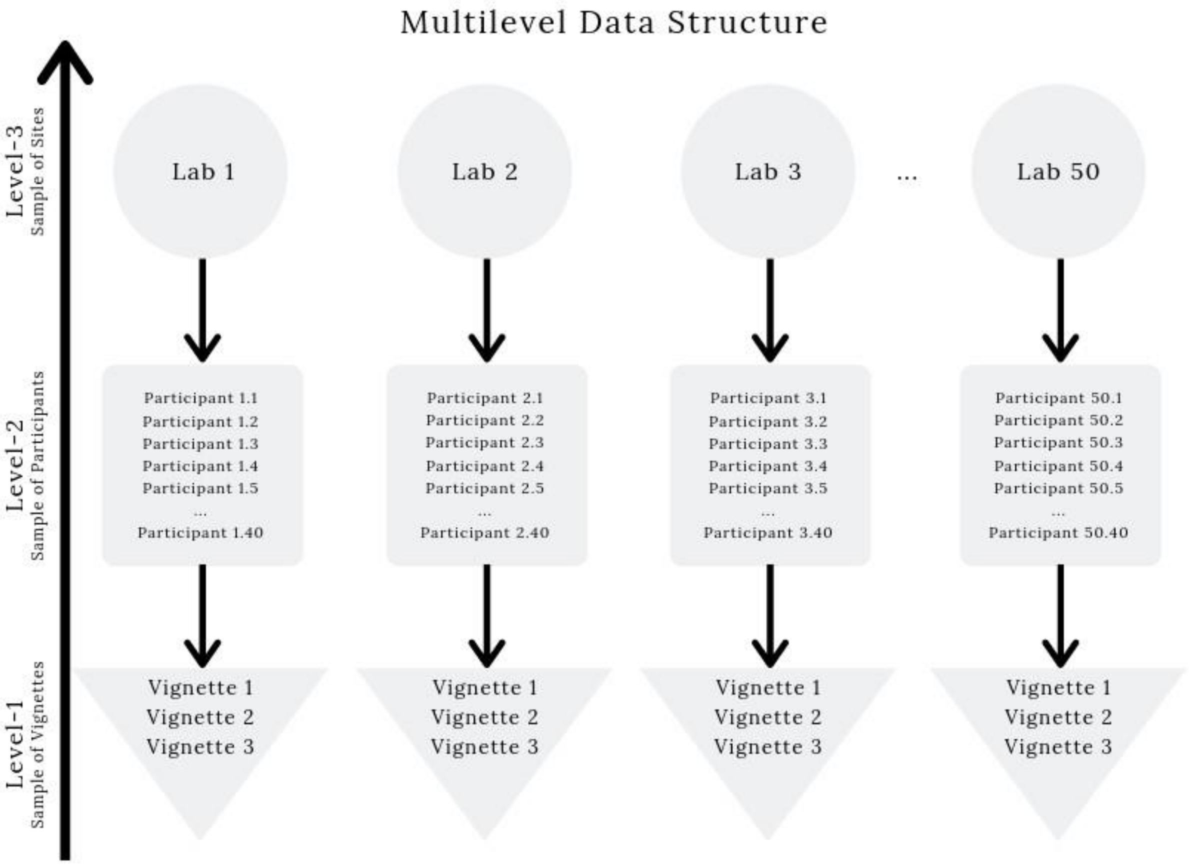


Figure 2: Data Structure. Total sample size includes a sample of labs (N = 50) each nested with a sample of participants (minimum N = 40) which are cross classified with a sample of three vignettes (stimuli) and three conditions. Each vignette is randomized and counterbalanced to one of three conditions (ignorance control, Gettier case, or knowledge control) to where each participant will be exposed to all three vignettes and all three conditions.

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 = .1) 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., 2015; Starman & 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 small real (between-subjects) effect of condition on knowledge attribution (β = .1). 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 measurements (i.e., vignettes) per participant, 32 participants per lab, and 9 labs (see Figure 3).

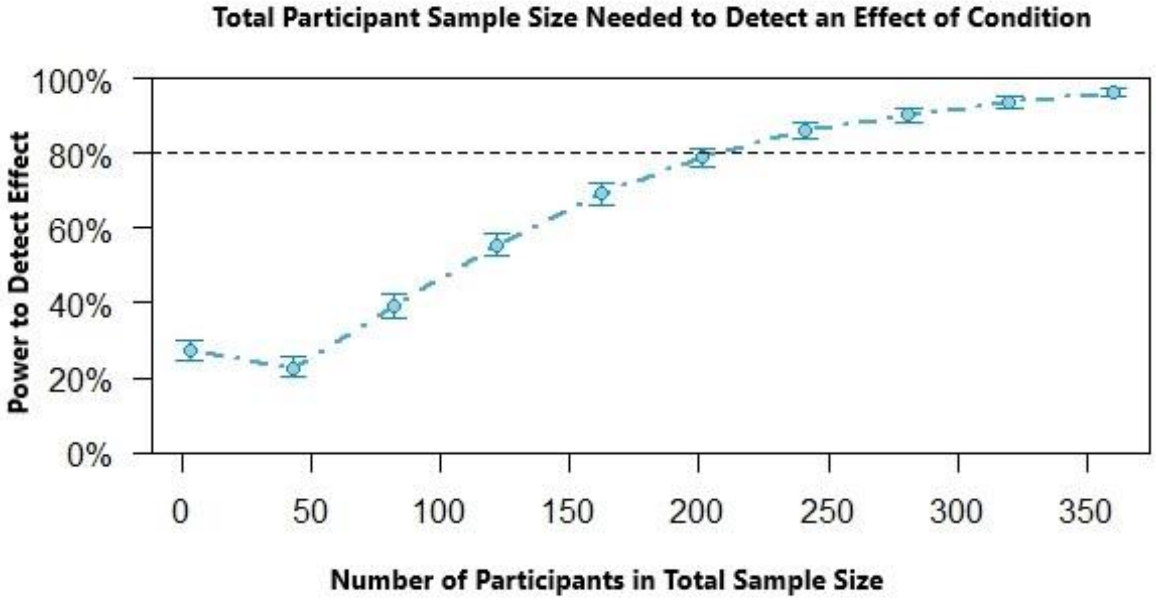


Figure 3: “powerCurve” (Green & MacLeod, 2016) plot for total number of participants needed to detect a small effect of condition on knowledge attribution (β = .1), with vignette count (*N* = 3) and lab count (*N* = 9) held constant, suggests we will achieve 90% power with a total participant sample size of 288 (864 total observations).

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* = 40) 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 49 sites signed up. The PSA currently has over 364 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. 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. 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 40 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 certification award for meeting the required sample size. To qualify for the CREP completion award for this study, a site must sample at least 40 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 and colleagues' Experiment 1 (which used a counterfeit object Gettier case), we decided to only sample other counterfeit object vignettes in order to test the generality of this class of Gettier cases, in lieu of testing Gettier cases more broadly. Thus, we first determined if a given vignette was a counterfeit object 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.

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] countries [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). For strictly educational purposes, contributors chose 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. 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[5], [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]. 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). 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

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

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. 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 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). 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 (country 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 countries (>20 countries) 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 country 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. 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 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.

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 4). 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.]

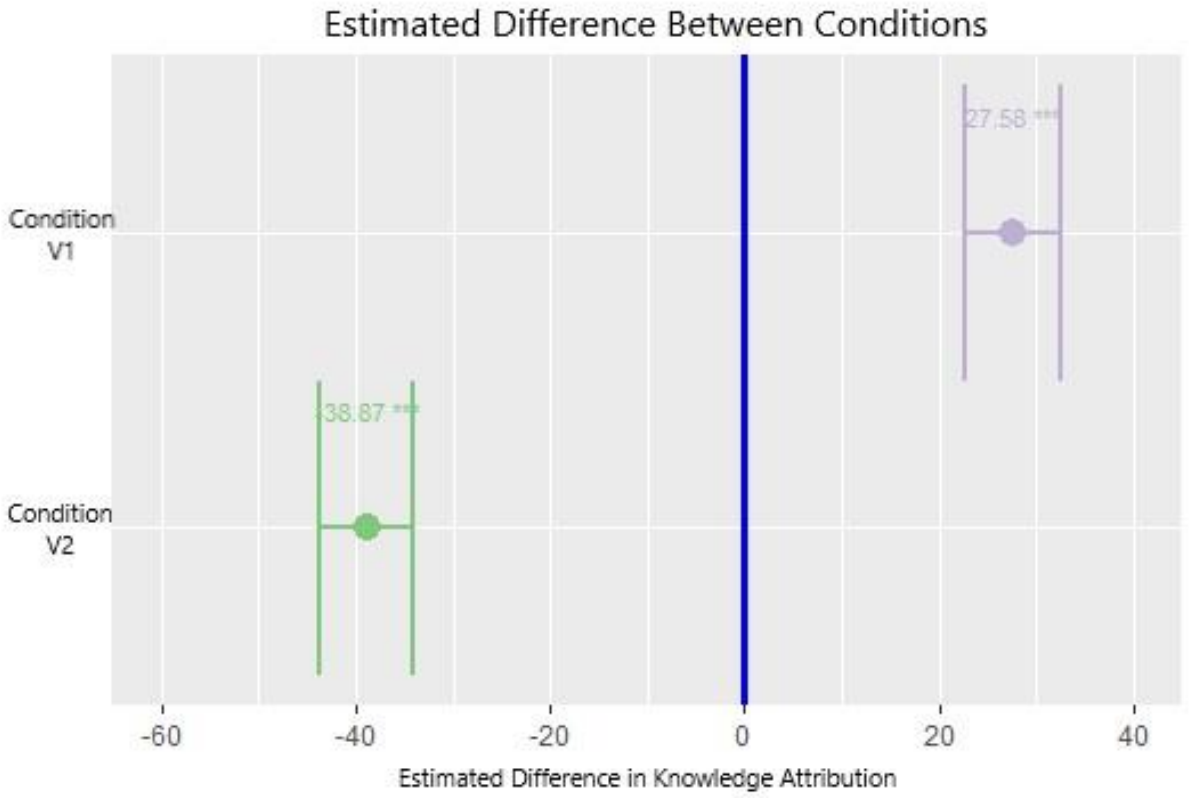


Figure 4. Example plot using simulation data to visualize the predicted difference between each condition, where Condition V1 is the estimated predicted difference between the Gettier case and the knowledge control (*b* = X.XX, *t*(XXX) = X.XX, *p* = .XX, 95% CI [X.XX, X.XX]) and Condition V2 is the 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.]

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.
I = Null (includes	X.X	X.X	X.X	X.X	X.X	X.X

condition)						
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): Country (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.

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).

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 5 for visualization of lab variation).

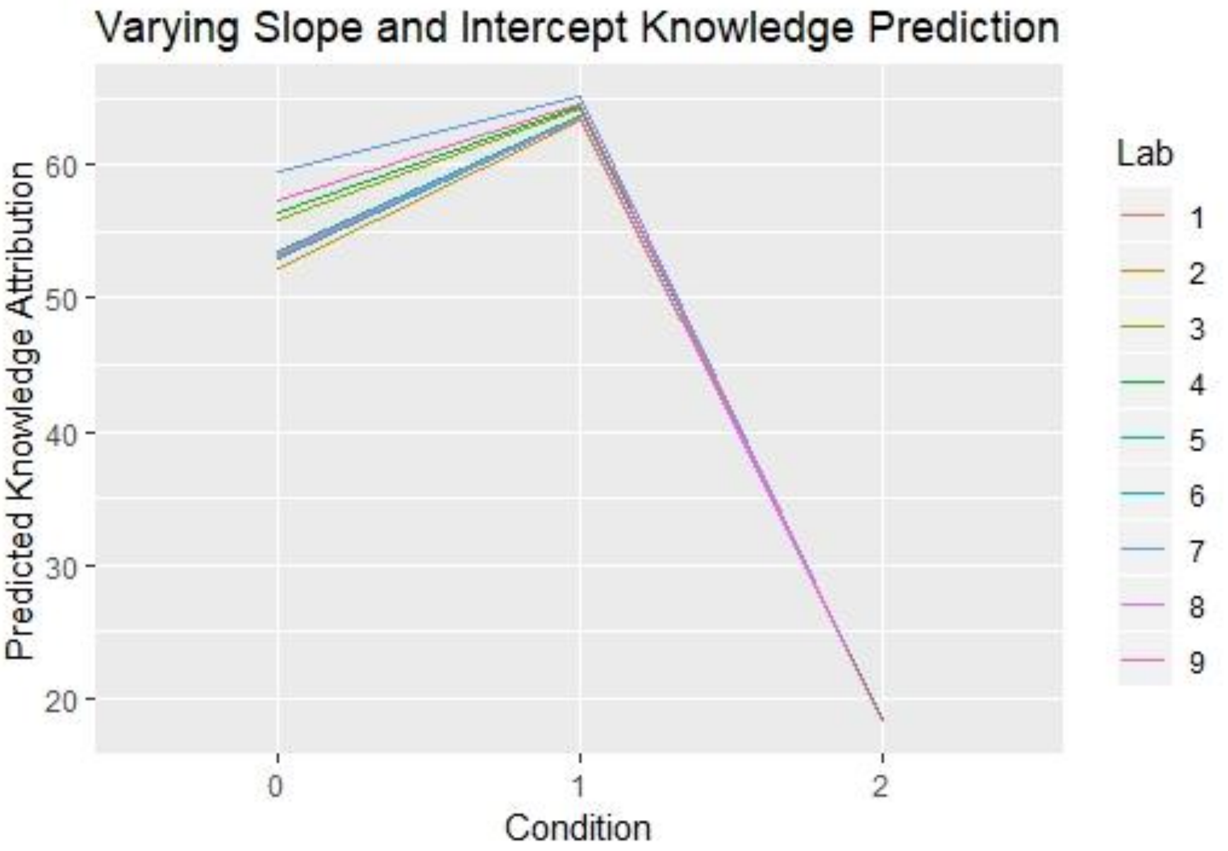


Figure 5. Example plot using simulation data to visualize the knowledge attribution multilevel model that allows for a random intercept and random slope for labs, where 0 = Gettier cases, 1 = knowledge controls, and 2 = ignorance controls. Actual data will be plotted based on the model of best fit.

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

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. 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). However, because most of our site samples likely lacked adequate power to detect the effects of covariates and were not very representative or balanced in regard to 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).

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

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): Country (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 aforementioned 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.

Discussion

X

Author Contributions: J.E.G. and C.R.C. conceived of the present study collaboration. B.F.H. drafted the manuscript and coordinated feedback and edits, M.J.B. and J.R.W. provided guidance with choosing the appropriate statistical approach, M.J.B., J.R.W., G.P., S.S., E.V., M.S., H.IJ., M.V., U.S.T., and A.F., provided guidance with refining the statistical and analytical models, H.IJ. conceived of and wrote the proposed data release plan, B.F.H. and J.R.W. wrote the R scripts for the planned analyses, B.F.H. simulated data and performed power analyses to estimate the required sample sizes to detect the main effects, J.R.W., C.R.C., G.P., S.S., E.V., E.D.M., L.D.C.A., S.A.S., M.S., A.F., E.K.H., H.IJ., M.F., A.A.O., M.R.A., S.C., W.M.S., A.T., A.T.N., T.R.E., M.V., U.S.T., S.T.H.E., D.R., S.O., N.M., F.P., A.H., M.J.M. F.C., J.D.A., K.M.W., M.J.B., V.D., G.K., C.S., Y.W., M.L., P.S., N.L., C.L., A.J.K., D.J., K.B., E.W., C.B., G.F., S.W., G.H., R.T.W., C.L., and G.H. provided feedback throughout manuscript construction on methodological decisions, study design, cultural differences between labs, improving clarity and undergraduate accessibility, addressing reviewer concerns, and copy editing. All authors have committed to collect samples for this study, except H.IJ., M.J.B., C.S., and Y.W.. All authors read and approved of the final manuscript. To determine author order, contributors were grouped into two categories: those who contributed to writing this manuscript and designing the present study (all authors mentioned in this subsection; Group A), and those who have only committed to collecting data (Group B). Based on the amount of contributions, the first six and the last three authors were given a set author order. All other authors were grouped and then both groups were independently randomized using the R script at the bottom of the primary analysis script, found on our OSF, with Group A then listed before Group B.

Conflicts of Interest: The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgments: The authors would like to thank all of the scientists on Twitter who provided suggestions and resources during the construction of this manuscript, particularly as it relates to our data simulations and analysis scripts. The authors would also like to thank Oscar Olvera for taking time out of their busy schedule to review our R script.

Funding: K.B. is supported by National Science Centre, Poland (2015/19/D/HS6/00641), E.V. is supported by the Swiss National Science Foundation (PZ00P1_154911 and PCEFP1_181141), M.J.B. is supported by the European Research Council (No 759320), and H. IJ. is supported by a French National Research Agency "Investissements d'avenir" program grant (ANR-15-IDEX-02).

Prior versions: Previous version of this manuscript are available at psyarxiv.com/zeux9/. Past versions contain most of the same content, with some corrections made to the design and analysis plan.

References

Behling, O., & Law, K. S. (2000). *Translating questionnaires and other research instruments: Problems and solutions.* Thousand Oaks, Calif: Sage Publications.

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.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F., Geller, J., Giner-Sorolla, R., … Veer, A. V. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology, 50*, 217–224. doi:10.1016/j.jesp.2013.10.005

Chartier, C. R., & McCarthy, R. J. (2018, March 28). StudySwap Tutorial. doi:10.31234/osf.io/wqhbj

Clark, H. (1973). The language as fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 335–359. doi:10.1016/S0022-5371(73) 80014-3

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

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

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

Gelman, A., & Hill, J. (2007). *Data Analysis using regression and multilevel/hierarchical models.* Cambridge, NY: Cambridge University Press.

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,* 493–498. doi:10.1111/2041-210X.12504

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

Hall, B. F., Grahe, J. E., Brandt, M. J., Chartier, C. R., Christopherson, C. D., Legate, N., … Wagge, J. R. (2018, April 8). Accelerated CREP -- Turri, Buckwalter, Blouw (2015). Retrieved from osf.io/n5b3w

Ichikawa, J. J., & Steup, M. (2018). The analysis of knowledge. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. Stanford, CA: Stanford University.

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

Kenny, D. A. (1985). Quantitative methods for social psychology. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology* (Vol. 1, 3rd ed., pp. 487–508). New York, NY: Random House.

Kenny, D.A., & Judd, C.M. (2019). *The unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication.* Unpublished manuscript.

Leighton, D. C., Legate, N., LePine, S., Anderson, S. F., & Grahe, J. E. (2018). Self-esteem, self-disclosure, self-expression, and connection on Facebook: A collaborative replication meta-analysis. *Psi Chi Journal of Psychological Research, 23*(2), 98-109. doi:10.24839/2325-7642.JN23.2.98

Nicols, S., Stich, S., & Weinberg, J. M. (2003). Meta-skepticism: Meditations on ethno-epistemology. In S. Luper (Ed.), *The skeptics* (pp. 227–248). Aldershot: Ashgate.

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

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

Machery, E., Stich, S., Rose, D., Chatterjee, A., Karasawa K., Struchiner, N., ... Hashimoto, T. (2015). Gettier across cultures. *NOÛS*, *51*, 1–20. doi:10.1111/nous.12110

Machery, E., Stich, S. P., Rose, D., Alai, M., Angelucci, A., Berniunas, R., ... Zhu, J. (2017). The Gettier intuition from South America to Asia. *Journal of the Indian Council of Philosophical Research*, *34*, 517–541. doi:10.1007/s40961-017-0113-y

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

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science, 1,* 501-515. doi: 10.1177/2515245918797607

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349* (6251), ac4716-aac4716. doi:10.1126/science.aac4716

Powell, D., Horne, Z., & Pinillos, N. Á. (2014). Semantic integration as a method for investigating concepts. In J. Beebe (Ed.), *Advances in experimental epistemology* (pp. 119–144). London, UK: Bloomsbury.

Powell, D., Horne, Z., Pinillos, N. Á., & Holyoak, K. J. (2015). A Bayesian framework for knowledge attribution: Evidence from semantic integration. *Cognition, 139,* 92–104. doi:10.1016/j.cognition.2015.03.002

Rosner, B. (2006). *Fundamentals of Biostatistics* (6th ed.). Duxbury: Thomson Brooks/Cole.

Satchell, L. (2018, August 7). [In progress] Collaborators for simple add-on “psych of psych” study. Retrieved from osf.io/ywekt

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

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., & Awtrey, E. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science,* 1, 337–356. doi: 10.1177/2515245917747646

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24,* 1875–1888. doi:10.1177/0956797613480366

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London, UK: Sage.

Sommet, N., & Morselli, D. (2017). Keep calm and learn multilevel logistic modeling: A simplified three-step procedure using stata, R, Mplus, and SPSS. *International Review of Social Psychology, 30,* 203–218. doi:10.5334/irsp.90

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

Starmans, C., & Friedman, O. (2012). The folk conception of knowledge. *Cognition, 124,* 272–283. doi: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,* 662–665. doi:10.1016/j.cognition.2013.05.009

Turri, J. (2013). A conspicuous art: Putting Gettier to the test. *Philosophers' Imprint, 13*, 1–16.

Turri, J. (2016a). Experimental epistemology and “gettier” cases. In S. Hetherington (Ed.), *Knowledge and the Gettier Problem* (pp. i-ii). Sydney, Australia: Cambridge University Press.

Turri, J. (2016b). Knowledge, certainty and assertion. *Philosophical Psychology, 29*, 293–299. doi:10.1080/09515089.2015.1065314

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

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

Vickstrom, E. R., Shin, H. B., Collazo, S. G., and Bauman, K.J. (2015). *How well — still good? Assessing the validity of the American community survey English-ability question*. Retrieved from the U.S. Census Bureau website: https://www.2020census.gov/content/dam/Census/library/working-papers/2015/demo/SEHSD-WP2015-18.pdf

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

Wells, G. L., & Windschitl, P. D. (1999). Stimulus sampling and social psychological experimentation. *Personality and Social Psychological Bulletin, 25,* 1115–1125. doi:10.1177/01461672992512005

Appendix A

Dummy Coding of Variables

Independent variable (IV_1):

- Vignette type, tested once for each participant, recorded in data output, dummy coded as:
    - ConditionV1
        - “Darrel” vignette = 0
        - “Gerald” vignette = 1
    - ConditionV2
        - “Darrel vignette = 0
        - “Emma” vignette = 1

Independent variable (IV_2):

- Reading condition, randomly assigned to participant, recorded in data output, dummy coded as:
    - ConditionV1
        - Gettier case = 0
        - Knowledge control = 1
    - ConditionV2
        - Gettier case = 0
        - Ignorance control = 1

Dependent variable (DV_1):

- Used as dependent variable in first multilevel linear regression analysis.
- Knowledge attribution, measured continuously from “only believes” to “knows”.

Dependent variable (DV_2):

- Used as dependent variable in second multilevel linear regression analysis.
- Reasonableness attribution, measured continuously from “unreasonable” to “reasonable”.

The following level-2 and level-3 variables will be included as predictors when constructing the random intercept model for each dependent variable (means and standard deviations will be calculated and reported for each).

Level-1 variables:

- Gender, measured as an open ended string response, will be grouped into three nominal categories (female, male, other). If we end up with enough participants who identify in the “other” category for a sufficiently powered comparison, then gender will be dummy coded as:
    - Gender1
        - Female = 0
        - Male = 1
    - Gender2
        - Female = 0
        - Other = 1
- Race/ethnicity, measured as an open ended string response, then grouped into similar nominal categories, dummy coded as:

	Race1	Race2	Race3	Race4	Race5	Race6	Race7

Black/African descent	1	0	0	0	0	0	0
White/European descent	0	0	0	0	0	0	0
Latin/Hispanic descent	0	1	0	0	0	0	0
Native American descent	0	0	1	0	0	0	0
Asian descent	0	0	0	1	0	0	0
Southeast Asian descent	0	0	0	0	1	0	0
Hawaiian/Pacific Islander descent	0	0	0	0	0	1	0
Native Australian descent	0	0	0	0	0	0	1

- Age, measured as an open-ended continuous response which can range from 0 to 120.

- Country of residence, measured as an open ended string response, and then grouped by country and dummy coded.
    - Countries will be coded accordingly:

	Region 1	Region 2	Region 3	Region 4	Region 5
Africa	1	0	0	0	0
Europe	0	1	0	0	0
Oceania	0	0	1	0	0
Asia	0	0	0	1	0
North America	0	0	0	0	0

Latin America/Caribbean	0	0	0	0	1

Level-2 variables:

- Test setting variables:
    - Setting
        - Measured dichotomously as whether they took the test in person or remotely, dummy coded as:
            - Took test in person = 0
            - Took test remotely = 1
    - Group size
        - For those tested in person, measured dichotomously as whether they were tested in a group or individually, dummy coded as:
            - Took test in a group = 0
            - Took test individually = 1

Appendix B

Vignette 1: Original "Darrel" Squirrel Vignette (Turri et al., 2015)

D1: "Darrel" Knowledge Control Condition

Darrel is an ecologist collecting data on red speckled ground squirrels in Canyon Falls national park. The park is divided into ten zones and today Darrel is working in Zone 3. While scanning the river valley with his binoculars, Darrel sees a 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."

Ecologists are unaware that a complex network of aquifers recently began drying up in parts of 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.

D2: "Darrel" Gettier Case Condition

Darrel is an ecologist collecting data on red speckled ground squirrels in Canyon Falls national park. The park is divided into ten zones and today Darrel is working in Zone 3. While scanning the river valley with his binoculars, Darrel sees a 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."

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 indeed a red speckled ground squirrel.

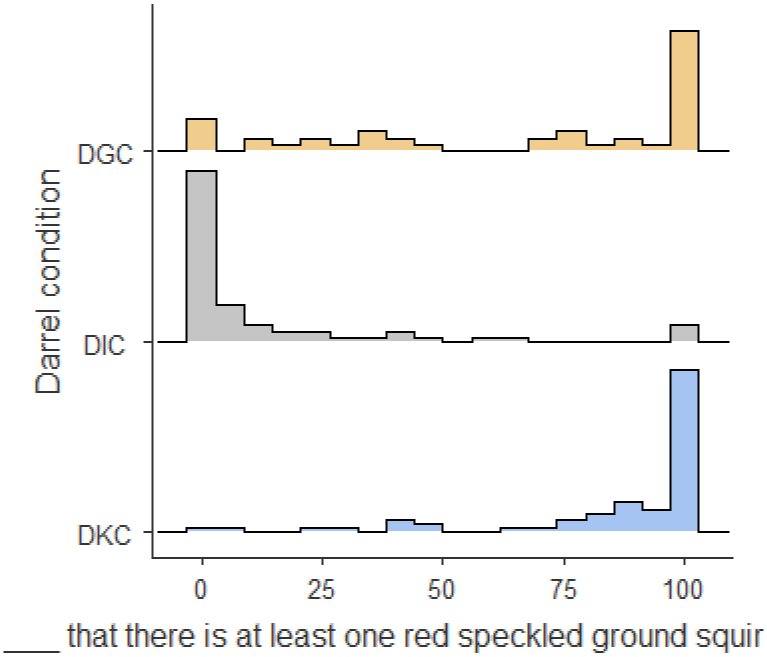
D3: "Darrel" Ignorance Control Condition

Darrel is an ecologist collecting data on red speckled ground squirrels in Canyon Falls national park. The park is divided into ten zones and today Darrel is working in Zone 3. While scanning the river valley with his binoculars, Darrel sees a 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."

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.

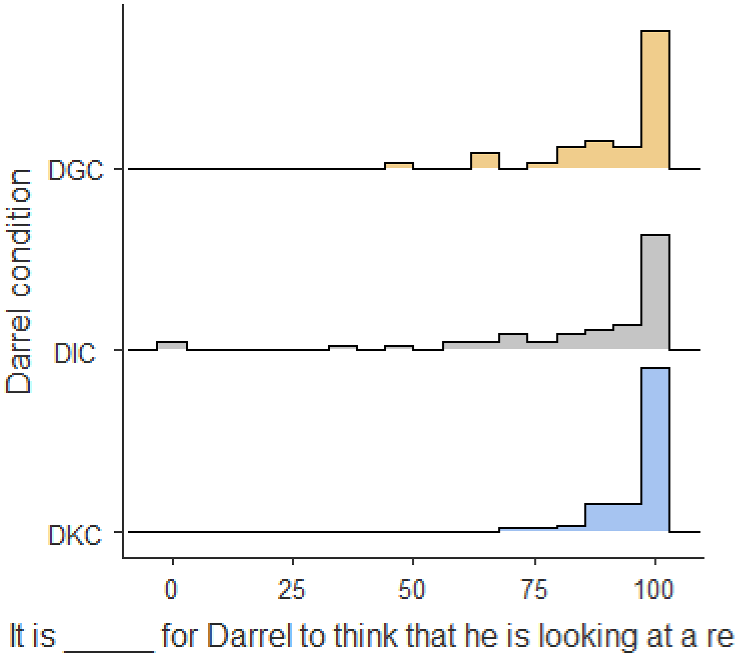
Questions (with bar charts demonstrating pretest response rates)

- ○ 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]

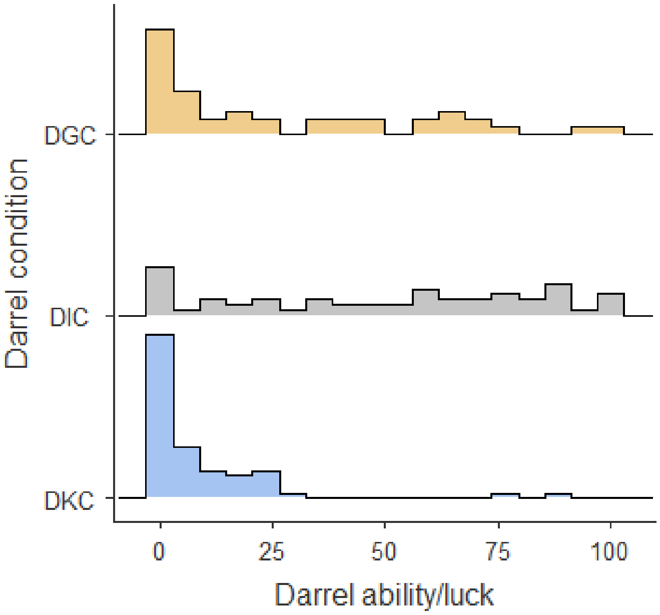


- ■ Darrel Gettier (*M* = 65.57, *SD* = 37.99)
- ■ Darrel ignorance (*M* = 13.49, *SD* = 25.49)
- ■ Darrel knowledge (*M* = 86.19, *SD* = 24.24)

- ○ 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:
        - ■ [(in)ability<------------------------> (good/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."]

Vignette 2: Modified[6] "Gerald" Fake Barn Vignette (Colaço et al., 2014):

G1: "Gerald" Knowledge Control Condition

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 rebuilt new houses in the place of the destroyed houses. These new houses look exactly like the old houses and can be used as actual housing.

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.

G2: "Gerald" Gettier Case Condition

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 fake houses in the place of destroyed houses. These fake houses look exactly like real houses from the road but are only for looks and cannot be used as actual housing.

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.

G3: "Gerald" Ignorance Control Condition

[6] The Gerald vignette was modified in two ways: First, we changed the words "house facade" to "fake house" to reduce the reading level required for comprehension; Second, we modified the structure of the vignettes to more closely match Turri et al. (2015) such that all conditions matched the qualities of belief present in the corresponding conditions from Turri et al. (2015) Experiment 1.

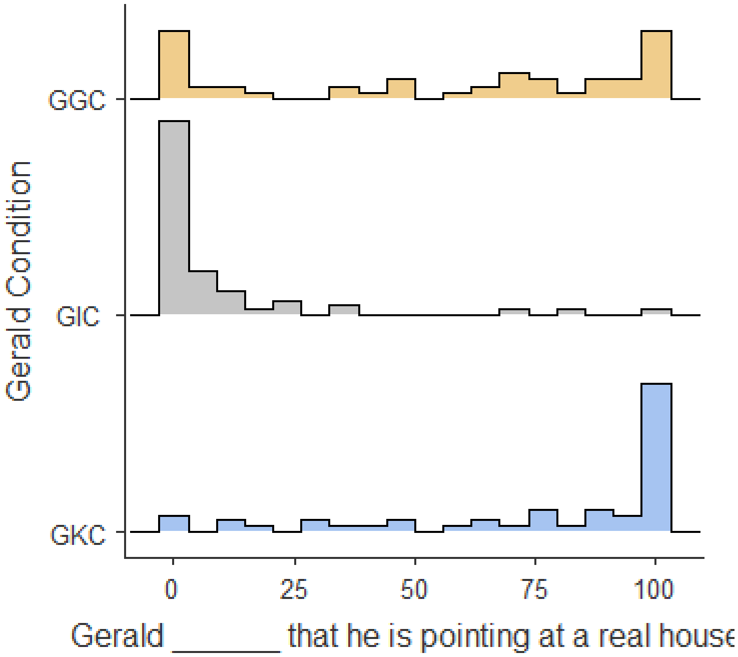
Gerald is driving through the countryside with his young son Andrew. Along the way he sees numerous objects and points them out to his son. ‘That’s a cow, Andrew,’ Gerald says, ‘and that over there is a house where farmers live.’ Gerald has no doubt about what the objects are.

What Gerald and Andrew do not realize is the area they are driving through was recently hit by a very serious tornado. This tornado did not harm any of the animals but did destroy most buildings. In an effort to maintain the rural area’s tourist industry, local townspeople built fake houses in the place of destroyed houses. These fake houses look exactly like real houses from the road but are only for looks and cannot be used as actual housing.

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

Questions (with bar charts demonstrating pretest response rates)

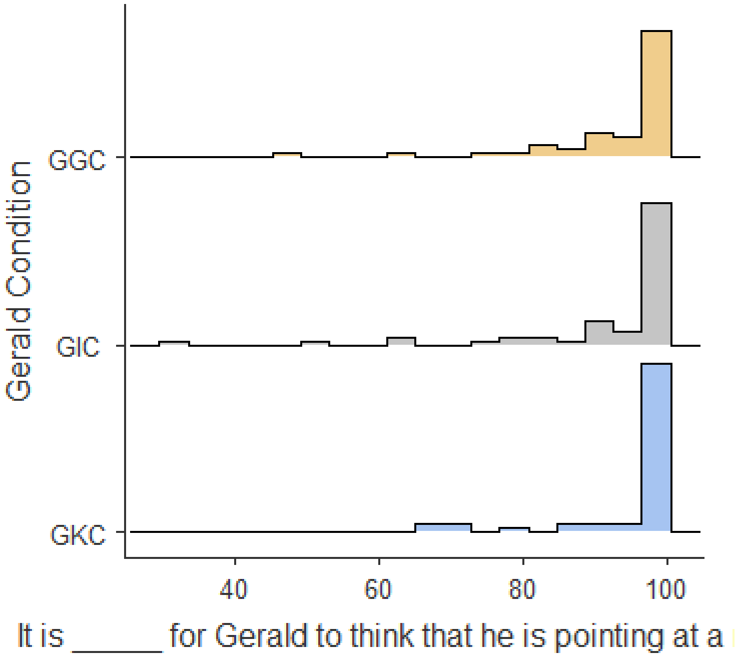
- ○ Primary knowledge probe (from Turri et al., 2015):
  - ■ **“Gerald ______ that he is pointing at a real house.”**
    - ● Visual analogue scale, 0-100:
      - ○ [only believes <------------------------> knows]



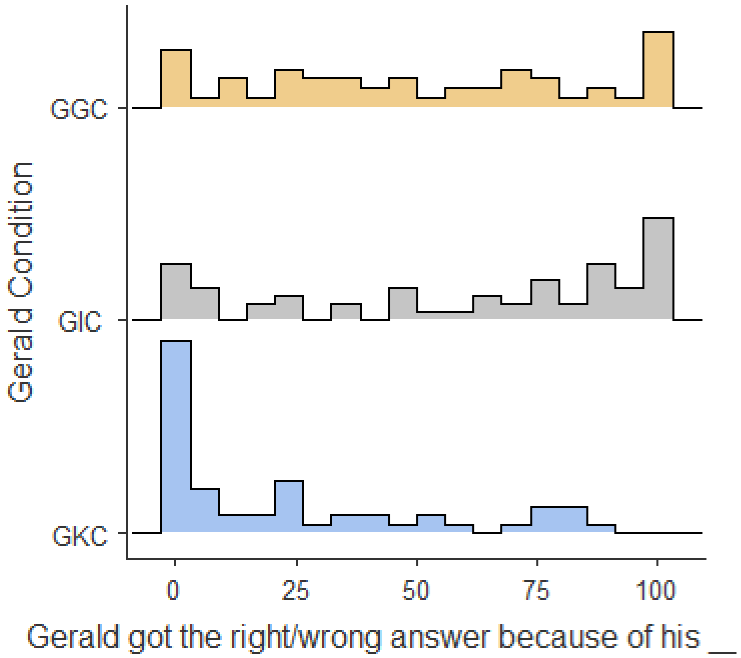
- ● Gerald Gettier (*M* = 58.28, *SD* = 38.68)
- ● Gerald ignorance (*M* = 8.35, *SD* = 19.35)
- ● Gerald knowledge (*M* = 76.96, *SD* = 32.29)

- ○ 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:
                - [(in)ability<------------------------> (good/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."]

Vignette 3: Modified[7] "Emma" Diamond Vignette (adapted from Nagel et al., 2013)

E1: "Emma" Knowledge Control Condition

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. However, this particular store has very honest employees and is known for their guaranteed authenticity; in the tray Emma chose, she picked out a diamond necklace from the selection of authentic diamond necklaces.

E2: "Emma" Gettier Case Condition

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 fact, this particular store has a very dishonest employee who has been stealing real diamonds and replacing them with fakes; in the tray Emma chose almost all of the pendants had cubic zirconium stones rather than diamonds (but the one she chose happened to be real).

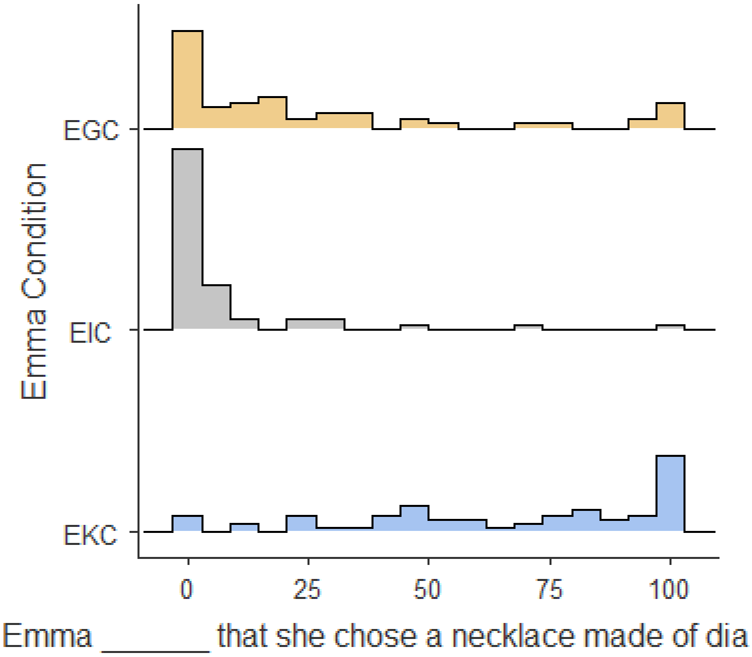
E3: "Emma" Ignorance Control Condition

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 fact, this particular store has a dishonest employee who has been stealing real diamonds and replacing them with fakes; in the tray Emma chose from, all of the necklaces – including the one she tried on – had cubic zirconium stones rather than diamonds.

Questions (with bar charts demonstrating pretest response rates)

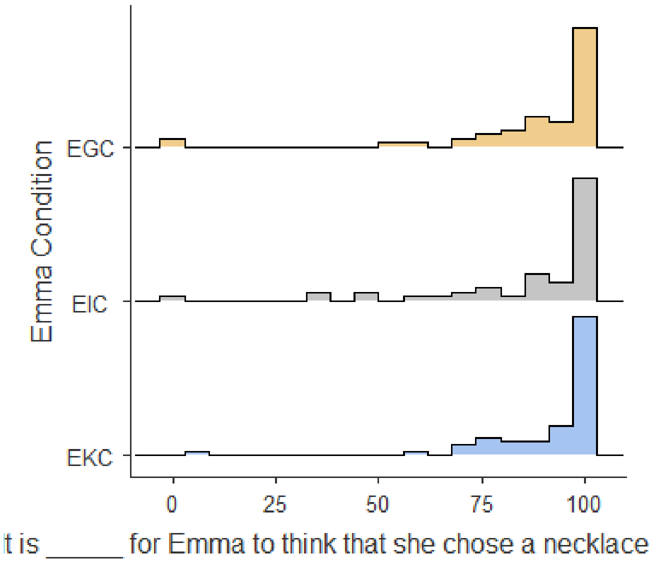
- ○ Primary knowledge probe (from Turri et al., 2015):
  - ■ **"Emma ______ that she chose a necklace made of diamonds."**
    - ● Visual analogue scale, 0-100:
      - ○ [only believes <------------------------> knows]

[7] For the "Emma" vignette, we changed the structure of the knowledge control condition to create more minimally matched pairs, while still maintaining the justified true belief described in it.

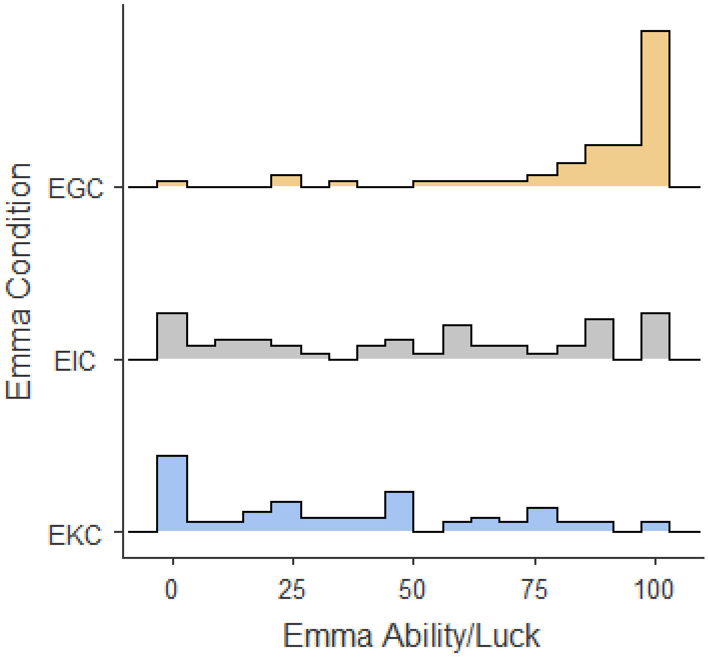


- Emma Gettier (*M* = 26.23, *SD* = 33.55)
- Emma ignorance (*M* = 8.44, *SD* = 19.26)
- Emma knowledge (*M* = 67.59, *SD* = 31.58)

- 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:
        - [(in)ability <------------------------> (good/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."]

Funneled Debrief Questions

- What do you think is the purpose of this study?
- What was your impression of the materials in this study?
- Have you ever participated in a similar study? If yes, please describe the study.

Appendix C

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: