PSCYOLOGISTS SUBSTANTIALLY (BUT INSUFFICIENTLY) UPDATE THEIR BELEIFS AFTER REPLICATION EVIDENCE

by

ALEXANDER DAVID MCDIARMID

ALEXA M. TULLETT, COMMITTEE CHAIR
MATTHEW R. CRIBBET
JAMES C. HAMILTON
WILLIAM P. HART
DEBRA M. MCCALLUM

A DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
in the Department of Psychology
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2021

ABSTRACT

The present research assessed if 1,096 psychologist participants sufficiently updated their beliefs in psychological effects when presented the results of multi-lab replication studies. In Phase I, participants read summaries of results from studies scheduled for replication attempts. For each study, participants made estimates of the population effect size and probability that the population effect was greater than d = .1 (i.e., non-trivial). During Phase I, participants were randomly assigned to a control or prediction condition with the only substantial difference being that those in the prediction condition were informed of the methodology for replication studies (not the results) of the original effects they evaluated and asked to predict how their confidence in the effect would change given various hypothetical replication study results. Approximately 1 to 1.5 years later, participants completed Phase II—the questions were the same for participants in the control and prediction conditions—in which they read summaries of replication results and provided revised effect size estimates and revised probability estimates. Participants' prior beliefs in original effects and replication evidence were quantified with Bayesian models which allowed us to model how a perfectly rational Bayesian agent would update their beliefs in original effects after incorporating replication evidence. While participants did update their beliefs substantially in the direction consistent with the replication evidence, as predicted, participants' confidence updates were insufficient for the weight of new evidence regardless of if confidence in psychological effects should have increased or decreased. Results suggest an impediment to scientific self-correction as it seems that psychologists underutilize replication evidence when updating their beliefs.

LIST OF ABBREVIATIONS AND SYMBOLS

b	Estimated values of raw (unstandardized) regression coefficients
CI (95%)	The range effect sizes that is wide enough for 95% of these ranges from different samples—from the same population and with the same sample size—would contain the population parameter
d	Cohen's measure of sample effect size for comparing two sample means
M	Arithmetic mean of the sample or subsample
N	Number of participants or cases in a sample
n	Number of participants or cases in a subsample
p	Probability associated with the occurrence under the null hypothesis of a value as extreme as or more extreme than the observed value
r	Estimate of Pearson product-moment correlation coefficient
SD	Standard deviation
SE	Standard error
t	Value expressing the difference between the mean of two subsamples
β	Estimated values of standardized regression coefficient

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my immense gratitude for all my friends, family, colleagues, and committee members who assisted me over the course of this research. It would have been all the more challenging to complete this research without your critical advice and support and your contributions tremendously improved the quality of the project. In particular, I would like to thank my undergraduate mentor, Dr. Monica Bartlett, and my graduate mentor, Dr. Alexa Tullett for their invaluable guidance. Finally, I would like to thank my fiancé Kristal Davis for her incredible support through the most challenging moments of this process.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

For science to be self-correcting, scientists must update their previously held beliefs when they become aware of new evidence. In light of a surprisingly high percentage of failed replications over the last decade, this may be particularly critical time period for psychologists to engage in appropriate belief updating (Open Science Collaboration, 2015; Ioannidis, 2005). The severity of the issue is perhaps best demonstrated by the results of Many Labs 1-3 as these replication studies are preregistered, solicit feedback from the original researchers, utilize large sample sizes, and are conducted by labs across the world for greater generalizability. Averaging the results of the three projects reveals an overall replication rate of 55% for 51 prominent findings (Klein et al., 2014; Ebersole et al., 2016; Klein et al, 2018).

Many contend that these replication findings (even if fair tests of the original effects) are nothing to be overly concerned about. False positives in the literature can stimulate future research that will, in turn, correct previous incorrect beliefs and advance our understanding of phenomena (e.g., Lieberman & Cunningham, 2009; Djulbegovic, & Hozo, 2007). In other words, there is a belief that science is self-correcting. For this to be an accurate depiction of the current state of psychology: 1) published replication studies must be commonplace, 2) Psychologists must read these replication studies, and 3) Psychologists must change their beliefs in response to this new evidence.

There has been a strong push for the publication of replication studies to be more prevalent in response to the "replicability crisis" (Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Frank & Saxe, 2012; Koole & Lakens, 2012). While published replications are arguably still too scarce, progress has certainly been

made in this regard. However, replication studies becoming more commonplace, and even psychologist routinely exposing themselves to replication results, is not enough by itself for psychology to be self-correcting. Psychologists must actually change their preexisting beliefs when they are in conflict with new evidence. But do they, and, if so, do they change their beliefs as much as is warranted by the new evidence?

Evidence directly addressing this issue is scant and tends to investigate biased evaluations of the quality of the evidence rather than the impact of the evidence on belief updating (or lack thereof). I intended to address this gap by investigating psychologists' belief updating in response to psychological evidence from (real) published original studies and replication studies. In doing so I aimed to start illuminating the degree to which psychology is, at present, a self-correcting science.

The present work is intended to more directly assess belief updating amongst psychologists by tracking their confidence in psychological effects before and after replication studies were run. The effects evaluated by participants were taken from two registered replication reports (replication studies that are peer reviewed before data collection; Mcarthy et al., 2018; Verschuere et al., 2018) and 25 studies from the Many Labs 2 project (ML2; Klein et al., 2018). In Phase I of the study, I measured psychologists' confidence in psychological findings after reading summaries of the original results. In Phase II of the study, I remeasured their confidence after they evaluated summaries of the replication studies for the effects they evaluated in Phase I.

Motivated Belief Updating

Given the extensive evidence demonstrating that people have the tendency to want to preserve their beliefs (e.g., Nickerson, 1998; Kunda, 1990**),** there is reason to think that

psychologists might be somewhat resistant to adequately updating their beliefs. While belief preservation is often aided by selective information exposure, even when people expose themselves to new evidence, they tend to be more critical of (and therefore less persuaded by) counter attitudinal information compared to pro attitudinal information (i.e., biased assimilation; Lord, Ross, & Lepper, 1979; Lord & Taylor, 2009; Miller, McHoskey, Bane, & Dowd, 1993; Nickerson, 1998**)**. Even people who consciously attempt to be open to changing their beliefs have difficulties succeeding at objectively assimilating new information as most of the mechanisms contributing to belief preservation occur outside awareness (for reviews, see Balcetis, 2008; Kunda, 1990).

Extensive evidence has also suggested that confirmation bias and other forms of motivated reasoning are intensified by personal investment in one's preexisting beliefs (Lord et al., 1979; Miller et al., 1993; Nickerson, 1998; Panagiotou & Ioannidis, 2012). In considering how this might apply to psychologists, it is notable that scientists may have particularly strong motives, relative to lay people, to reach accurate conclusions; learning the truth about phenomena is the inherent goal of science. However, if psychologists are personally invested in their skepticism or confidence in psychology findings (which is perhaps fairly likely for findings in their area of expertise) they would also have strong directional motives that could limit their belief updating. When replication results conflict with preexisting beliefs, psychologists may have directional motives that compete with their accuracy motives.

The tendency to engage in motivated reasoning has been observed among scientists despite their accuracy motives. They are more critical of manuscripts and abstracts with uncongenial findings (Hergovich et al., 2010; Mahoney, 1977); judge studies to be of greater

quality when they support their preexisting views (Koehler, 1993); and interpret data in ways that are consistent with preexisting views (Panagiotou & Ioannidis, 2012). However, it remains remains unclear if these tendencies extend to psychologists' belief updating for original findings. While previous research provides substantial evidence that scientists are critical of novel research methods when the results are uncongenial (Hergovich et al., 2010; Mahoney, 1977; Koehler, 1993; Panagiotou & Ioannidis, 2012) these trends may or may not translate to evaluating replications. Furthermore, previous research has focused primarily on quality evaluations of studies' methodology and did not track changes in beliefs about the truth of effects after exposure to the new evidence. The present work aims to address these gaps in the literature by tracking psychologists' belief updating after they view new evidence from replication studies.

Motivated belief updating among scientists might also be the result of motivations to achieve belief closure rather than motivation to reach a particular conclusion. Under such conditions, once establishing an initial belief ("seizing") people tend to "freeze" on this original belief rather than updating the belief based on new information **(**Krunglanski & Webster, 1996**)**. Doing so allows one to remain in a state of cognitive closure rather than expending limited cognitive resources on forming (or considering forming) a new belief. If such a process is involved in scientific belief updating for psychologists, once achieving a degree of nonspecific (in that no *particular* conclusion was desired) cognitive closure regarding a scientific effect, psychologists might have the tendency to discount evidence that would otherwise cause them to reconsider their original conclusion.

To assess the role of motivated reasoning in assimilating new evidence from replication studies, the present research assessed how belief updating is related to personal investment in the

original finding. Motivated reasoning was also assessed by considering if psychologists believe it would be appropriate to update their beliefs in response to hypothetical evidence to a greater extent than they are actually willing to truly update their beliefs when assimilating evidence from real replication studies. Third, the role of motivation reasoning in assimilating replication evidence was also assessed by measuring if psychologists are less critical of replication studies' methodology *before* learning the results of the studies relative to their criticism of replications' methodology *after* learning that the results of a replication conflict with their preexisting beliefs.

Individual Differences and Motivated Reasoning

It is intuitively appealing to suspect that psychologists might be more successful than lay people in avoiding motivated reasoning – at least for uncongenial psychological data. Surprisingly, however, research has observed various indicators of analytical skills (e.g., objective knowledge of subject matter, SAT scores, science literacy, and numeracy) to be either unassociated with self-serving evaluations of evidence (Ballarini & Sloman, 2017; Taber et al. 2009; Stanovich & West, 2008), or *positively* associated with self-serving evaluations of evidence (Kahan et al., 2017; Kahan et al., 2012; Taber, Cann, & Kucsova, 2008; Taber & Lodge, 2006). In the latter case, this evidence suggests that in some contexts people who are most capable of intelligently critiquing flawed evidence utilize these skills to identify flaws with uncongenial and neutral evidence but not to identify flaws with congenial evidence. In considering the status quo for scientific self-correction, it is particularly noteworthy that greater scientific literacy is associated with greater biased assimilation of new data regarding political issues (Kahan et al., 2012).

Kahan and colleagues (2017) have described this phenomenon as motivated numeracy and note that highly numerate lay people only utilize their numeracy when doing so does not threaten their cultural identity. While it is not cultural identity, precisely, that is at stake for psychologists evaluating new data, they are nonetheless often personally invested in their conviction in or skepticism towards particular findings. It is therefore consistent with the aforementioned findings to suspect that psychologists may be similarly motivated to selectively utilize their skills when updating their beliefs in response to new data. For example, expert knowledge in understanding barriers to belief updating might not actually help psychologists to think and behave in a manner to overcome barriers to belief change. The present research considered if skills and traits that would intuitively seem advantageous for unbiased belief updating (professional expertise in understating the influence of motivated bias and intellectual humility) are actually associated with greater belief updating when greater belief updating is warranted by the new evidence.

Unmotivated Belief Updating

Alternatively to motivated reasoning, it may also be the case that while psychologists are susceptible to confirmation bias this could be the result of cognitive bias (rather than motivated bias). Instead, extensive research suggests that under-adjusting to new information may simply be due to the general tendency of people to be more affected by the information that they assimilated first (Bruner & Potter, 1964; Green & Donahue, 2011; Jones, Rock, Shaver, Goethals, & Ward, 1968; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012; Schul & Burnstein, 1985**)**. If an unmotivated account explains psychologists' belief updating, then we should expect that while they may still be under-influenced by new evidence, their lack of updating would be not related or minimally related to their degree of personal investment in the original finding.

Minimal belief updating could also not be due to either motivated or cognitive biases. It would not be indicative of bias for psychologists to not update their beliefs to accommodate new findings when, for instance, they are critical of the quality of a replication study even before knowing the results. Replications are frequently criticized for procedural differences that harm the replication quality and for contextual differences (i.e., hidden moderators) from the original study (Van Bavel, 2016). In cases when psychologists reach these types of subjective judgments, more restricted belief updating is more rational than the larger normative adjustments that are appropriate for close replications (i.e., replications that are as close to the original study as is reasonably possible). However, these more limited belief updates are more justified if shortcomings in the replication are identified before the results are known. If psychologists are only critical of replication studies after seeing results that conflict with their preexisting belief, this would imply that motivated reasoning is leading to restricted belief updating; the replication study did not threaten their directional conclusion motives before the results were know. In the present work, I will investigate how belief updating is associated with changes in perceptions of the quality of replication studies' procedures after the results are known

Additionally, many psychologists' prior beliefs are based on a large amount of evidence beyond a single original finding. From a Bayesian perspective, people should incorporate this additional information into their prior and a greater amount of new evidence would be necessary for a large belief update to be rational. Therefore, it could appear from an analysis of just the evidence from an original study and contradictory replication that a scientist is under-updating when, in reality, their prior is simply stronger than what would be justified by the original study alone. In the present work, when determining how much psychologists

should update their beliefs (according to a Bayesian model), we addressed this consideration by asking participants specify their own priors' beliefs and their certainty about those beliefs. Thus, if participants start out with a very confident prior belief, our normative model will indicate they should update less.

Overview of Current Study

The present work assessed if psychologists do in fact update their beliefs in psychological findings in response to evidence from replication studies and, if so, if they update their beliefs sufficiently given the weight of the new evidence (see Bayesian Updating section for an explanation of basic Bayesian concepts and our strategy for assessing "sufficient" updating). Second, I assessed if the degree to which psychologists adjust their beliefs might be attributable to motivated reasoning (specifically biased assimilation) serving the function of maintaining preexisting beliefs. Third, I assessed if individual difference variables (specialized knowledge of motivated biases and intellectual humility) moderate the extent to which participants update their beliefs.

The study was conducted in two Phases spaced approximately 1 year apart with participants randomly assigned to a prediction condition or control condition. During Phase I, participants in both conditions evaluated three (real) original studies. These evaluations involved estimating the effect size in the population and estimating the probability that the effect size was *substantially* greater than zero in the population (details in Methods section). Participants did not yet know they would later be shown replication evidence when evaluating the original studies. Participants in the prediction condition only then read about the methodology of (real) replication studies attempting to replicate the original effects they had evaluated and predicted how their evaluations of the original studies' effects would change given various hypothetical

replication results. Participants in the control condition did not see any details about the replication studies until Phase II—though they were asked if they already knew the results of any study attempting to replicate the original effects they evaluated. For each study participants in both conditions also reported (during Phase I) how personally invested they were that future research would support the effect and participants in both conditions evaluated the merits of the replication study methodology (prediction condition: Phase 1 [before results were know] and Phase II [after results were know]; control condition: Phase II only). Last, participants answered items measuring individual differences: intellectual humility, and profession expertise in confirmation bias and motivated reasoning.

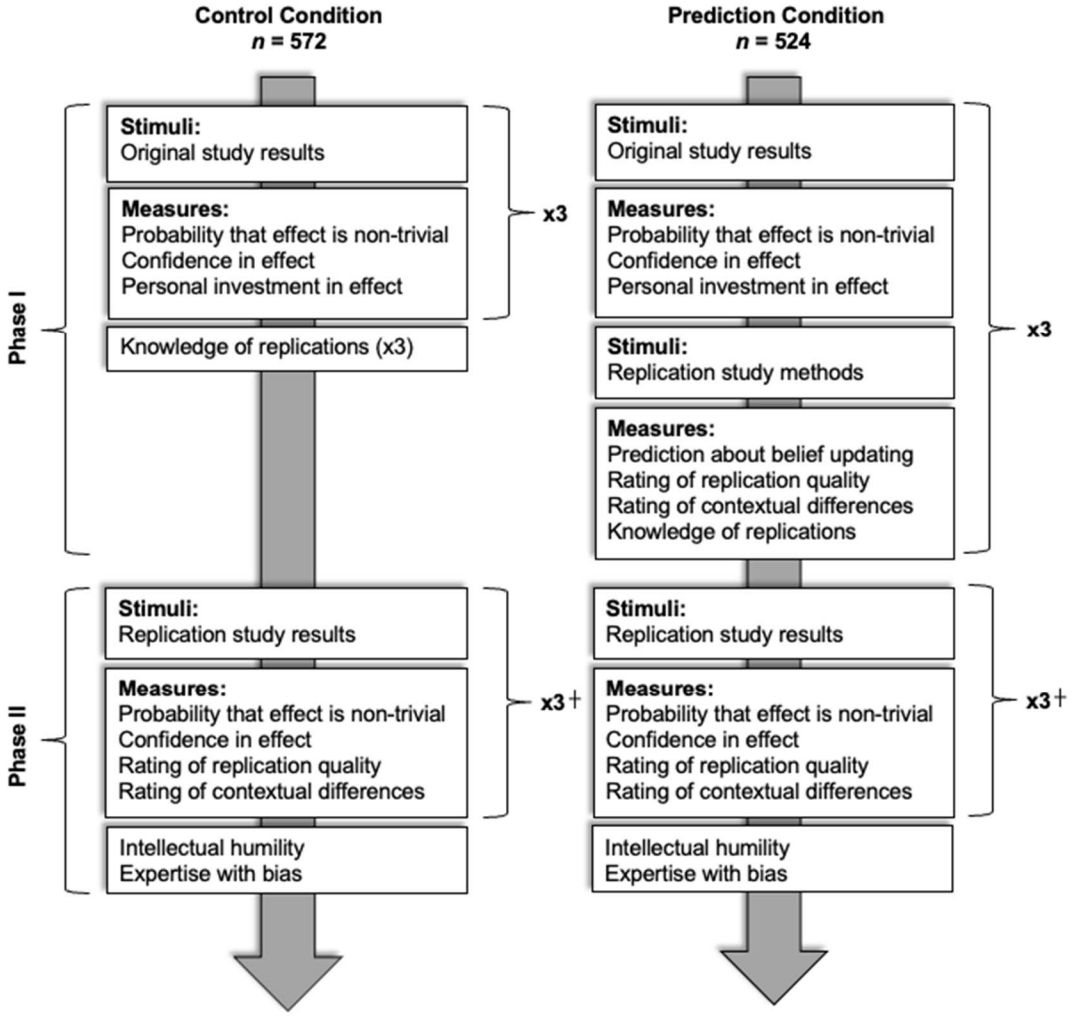
METHODS

My primary aim with this research was testing to what extent psychologists update their beliefs in response to new replication evidence and to assess if their belief updating was as strong as belief updating warranted by the strength of the replication evidence. Secondary aims were to test for evidence of motivated reasoning influencing belief updating and appraisals of replication study methodology, and to test if individual differences moderated belief updating. For hypotheses, see the ANALYSIS PLAN chapter. Study materials, data, preregistration, and preregistration addendums are all available on the Open Science Framework (https://osf.io/fmk6y/?view_only=b84feec005344530a0c74b76e4c3916f). See Figure 1 for an overview of study procedure.

Design

The design was a 2 x 2 mixed design with Time (Phase I vs. Phase II) as a within-subjects variable and Condition (Prediction vs. Control) as a between-subjects variable. The Control condition was included so I could assess the possibility that the mere act of making predictions about belief updating affects subsequent belief updating. If making predictions does serve as an unintended manipulation, I wanted to be able to control for this effect to better measure psychologists' belief updating in a more ecologically valid context. The Prediction condition served the function of comparing hypothetical belief updating—which would tend to involve less motivated reasoning as the task does not, in fact, invite participants to report true

Figure 1. Overview of Study Procedure



+Some participants saw fewer than three studies during Phase II because some studies were dropped (see main text for details)

(non-hypothetical) belief updating—with true belief updating in response to real replication evidence.

Stimuli

In Phase I, participants evaluated 3 out of a pool of 35 social psychology studies. These 35 studies were the targets of large-scale replications – for which the results had not yet been published – from the Many Labs 2 project (n = 23; Klein et al., 2018) the

Many Labs 5 project (n = 9; Ebersole, et al., 2019), and two other registered replication reports (n = 3; Mcarthy et al., 2018; Verschuere et al., 2018). Due to it being impractical to delay Phase II until the Many Labs 5 project and one of the registered replication reports were published, these studies were not included for ratings in Phase II. Because in Phase II participants only revaluated the effects they had seen in Phase I, this meant that some participants evaluated fewer than 3 studies in Phase II.

Participants

One-thousand eight-hundred and twenty psychologists completed Phase I of the study. The only eligibility requirement was that participants were graduate students or more senior (e.g., post-doc, faculty member), in the field of psychology. Participants were recruited from psychology conferences and professional psychology listservs. Participants were compensated with a $10 Amazon gift card for completing Phase I and a $20 Amazon gift card for completing Phase II.

Of the1,820 participants, 543 were lost to attrition due to not accepting their invitation to complete Phase II (n = 482), not giving a valid email for recontact (n = 7), or completing Phase I multiple times (n = 54). An additional three participants could not complete Phase II because none of the studies they viewed in Phase I were included in Phase II. Of the remaining 1,274 participants, an additional 178 were excluded from analyses because, they evaluated considerably more than three studies due to experimenter error (n = 14), were undergraduate students (n = 25), or indicated they did not understand critical questions (n = 139). I did not preregister our exclusions for experimenter error. The latter two exclusion criteria were: 1) preregistered as exclusion criteria (did not understand the critical questions), and, 2) specified in

the consent form as disqualifying for study eligibility (undergraduate students). This left me with responses for 1-3 studies from 1,096 participants.

Phase I Procedure

For participants recruited from conferences, the research team introduced themselves on the conference premises to conference attendees (with a conference badge) and informed them of the purpose of the study. Participants were informed of the financial compensation for the study and that they were being asked to participate in a two-phase study in which they would be reporting on their beliefs about contemporary research findings. Before consenting to participate, participants also verbally confirmed that they met the study's eligibility criteria for the study: completed at least some graduate school level training in psychology. Participants recruited from professional psychology list servs were provided a link to participate in the study online within the advertisement for the study. The advertisement informed participants of the study's: purpose, financial compensation, and eligibility criteria.

After providing informed consent to participate and their email address at which to be contacted for Phase II, participants began by reading a short summary of an original study (see Appendix A for an example) with the key finding underlined for clarity. For each study participants were informed of the study authors, the journal that published the study, the sample size, the p-value, the effect size, the 95% confidence interval for the effect size, and, when applicable, group means. Participants were then provided an explanation of Cohen's guidelines for describing effect sizes and told the average effect size in social psychology. Next, participants estimated the probability that the study's key effect was trivially small. This estimate constituted the participant's *Prior* (See analysis

plan for precise definition of *"Prior*" and use in hypothesis testing). I clarified that I was defining "trivially small" as a Cohen's *d* of less than .1 and that they were estimating this probability for the effect size in the population (i.e., the parameter) as opposed to in the sample. Participants also estimated the effect size in the population. Using their *Prior* and effect size estimate I was able to model a density distribution expressing the relative probabilities participants estimated for different potential effect sizes in the population (see Calculating Prior Distributions and Bayesian Posteriors section for details). Participants were also asked to report on, "to what degree are you invested in this finding" using a 5-point Likert scale (1 = strong preference that it be refuted, 3 = no preference, 5 = strong preference that it be supported). All participants, regardless of condition, responded to these measures for three studies.

At this point, the survey procedure diverged between conditions. Participants in the control condition moved on to evaluate their second study whereas those in the prediction condition completed more measures pertaining to the first study they evaluated. Participants in the prediction condition were then shown a summary of the replication study's protocol (see Appendix A for an example) reminding them of the key finding from the original study, highlighting all known differences from the original study, and describing the sample for the replication study including the anticipated sample size. They then rated the quality of the replication study using a 3-point scale (*1 = low quality, 2 = moderate quality, 3 = high quality*), and the extent of contextual differences between the replication and original studies using a 3-point scale (*1 = minor contextual differences, 2 = moderate contextual differences, 3 = major contextual differences*).

Next, participants in the prediction condition were asked to consider six hypothetical scenarios of potential effect sizes (*d* = -.15 [opposite direction of the original finding], .05, .20,

.60, 1.2, 1.8) that might be found by the replication. The actual results of the replication—which I already had access to but were not yet available to the public—all fell within this d = -.15 to 1.8 range. Participants were also reminded of the *Prior* (probability that the population parameter in greater than trivial) and population parameter estimate they had specified based the original study. For each scenario participants predicted what they would estimate for the population parameter if the hypothetical replication results were real. These estimates provided by the participants were used to interpolate their predictions for any prospective replication results (see Analysis Plan section for details). Following responses to all three studies participants in both conditions concluded Phase I by providing demographic information and responding to a measure assessing their self-perception of their success in comprehending questions.

Phase II Procedure

Participants completed Phase II roughly 1 to 1.5 years after they had completed Phase I (after the replication results were made public). Regardless of the method for Phase I recruitment, participants who complete Phase I were recontacted via email to invite them to participate in Phase II. Participants were reminded of the purpose of the study and financial compensation for completing Phase II. Participants were offered at least three weeks to complete Phase II from the day they received their Phase II invitation, but some participants had up to 9 weeks to complete Phase II. Phase II invitations were staggered to mitigate the risk of errors in financially compensating participants.

The Phase II procedure was the same across conditions. Participants were first reminded of the original study results with the same summary they read in Phase I—to

mitigate the risk of demand characteristics and enhance ecological validity, they were not reminded of their estimates for the original study. Participants then read a summary of the replication protocol and results investigating the same effect as the original study. These replication results summaries reminded the participants of known differences from the original study and were otherwise formatted identically to the original study summaries (i.e., provided sample size, effect size, etc.). Participants then evaluated the replication study by rating the quality of the replication study and reporting on perceived contextual differences between the original and replication studies. For participants in the prediction condition, responses to these measures could be compared to their responses to the same measures in Phase I. This enabled me to measure if their critiques of the replication study changed after viewing the results.

After participants evaluated all three of their assigned replication studies, they completed the Leary and colleagues (2016) measure of intellectual humility. Participants also completed a measure assessing their perceptions of their professional expertise in both confirmation bias and motivated reasoning. Responding to two 5-point Likert scales items, participants were asked to rate their expertise in both psychological concepts compared to the expertise of the average psychologist (1 = *Much less* than the average psychologist, 2 = *Slightly less*, 3 = *About the same*, 4 = *Slightly more*, 5 = *Much more*). The ratings were averaged to compute the *Bias Expertise* variable. Participants then provided demographic information including age, gender, race, political ideology, career position and subfield within psychology. To conclude the study, participants were presented the study debriefing.

ANALYSIS PLAN

See Table 1 for an abbreviated description of hypotheses (see Hypotheses section for full details).

Calculating Prior Distributions and Bayesian Posteriors

I converted each individual's responses about their estimates of the effect population parameter and estimate that the effect population parameter is Cohen's $d > .1$ (we termed the later estimate participants *Actual Prior*) into probability density distributions (i.e., their *Prior Distribution*). To do this, I assumed these distributions were normally distributed. I used the participant's Phase I estimate of the population effect size as the mean of the distribution, and their *Actual Prior*, to calculate the standard deviation of the modeled *Prior Distribution*. For illustration, consider a participant who guesses that the population d = .25 and there is a 32% chance d < .1. This establishes that 18% of the distribution is between $d = .1$ and $d = .25$. We also know that 18.3% of a normal distribution is contained between the mean and one standard deviation below the mean. Therefore, we know that such a participant has distribution with a standard deviation of (approximately) .15 (.25 - .1).

In cases where participants' estimates were incompatible with the assumption of a normal distribution (e.g., a distribution mean of .1 with 90% of the distribution greater than .1), we did not calculate *Bayesian posteriors* (and thus these participants were excluded from analyses involving *Bayesian posteriors*—this exclusion was not preregistered). Because prior distributions cannot be defined for *actual priors* of 0% or 100%, we converted responses of "0%" and "100%" to .25% and 99.75%, respectively. While my assumption that participants' prior effect

Table 1. Hypotheses

	Prediction when participants should, according to our Bayesian model, adjust downward	*Prediction when participants should, according to our Bayesian model, adjust upward*
RQ1: How much psychologists update their beliefs in response to new empirical evidence?		
H1. Participants will update their beliefs in response to new evidence	*Actual posteriors* will be lower than *actual priors*	*Actual posteriors* will be higher than *actual priors*
H2. Participants will not update as much as our Bayesian model would predict	*Actual posteriors* will be higher than *Bayesian posteriors*	*Actual posteriors* will be lower than *Bayesian posteriors*
H3. (Prediction condition only) Participants will not update their beliefs as much as they predict they will	*Actual posteriors* will be higher than *predicted posteriors*	*Actual posteriors* will be lower than *predicted posteriors*
RQ2: Do psychologists show evidence of motivated reasoning when evaluating replications?		
H4. (Prediction condition only) Participants will provide lower ratings of replication quality when this would preserve their pre-existing beliefs	Higher *actual priors* will be associated with greater decreases in ratings of replication quality from Phase I to Phase II	No prediction
H5. (Prediction condition only) Participants will be more likely to identify context differences when this would preserve their pre-existing beliefs	Higher *actual priors* will be associated with greater increases in context difference ratings from Phase I to Phase II	No prediction
RQ3: What predicts the extent of psychologists' belief updating?		
H6. Personal investment will be negatively associated with belief updating	Participants who are more personally invested in a finding will show less of a difference between their *actual priors* and their *actual posteriors*	No prediction
H7. Expertise regarding confirmation bias and motivated reasoning will be positively associated with belief updating	Participants who have more expertise regarding confirmation bias and motivated reasoning will show more of a difference between their *actual priors* and their *actual posteriors*	
H8. Self-reported intellectual humility will be positively associated with belief updating	Participants who self-report more intellectual humility (Leary et al., 2017) will show more of a difference between their *actual priors* and their *actual posteriors*	

size beliefs are normally distributed seems likely to approximately model participants' true beliefs, this assumption is nonetheless imprecise. For example, if participants systematically had

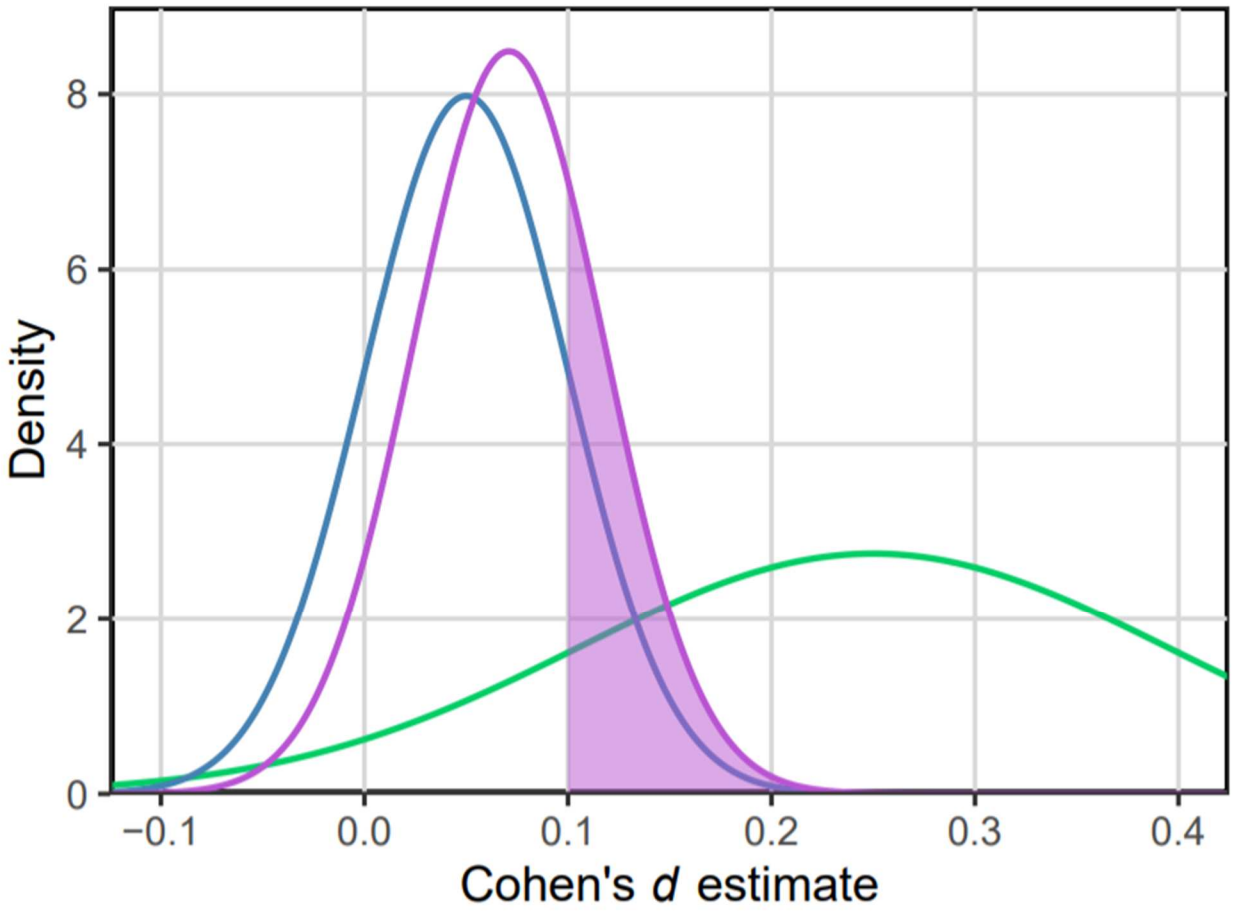
more leptokurtic [platykurtic] bell shaped distributions, our model would systematically overestimate [underestimate] how much participants should adjust.

For each *Actual Prior*, I calculated a *Bayesian Posterior* based on the replication evidence from Phase II, using conjugate Bayesian inference. That is, I considered how a perfectly rational Bayesian agent would update their beliefs based on new evidence, starting from participants' actual priors, and conditional on them trusting the replication results. (see formulas and spreadsheet to test different inputs here: https://osf.io/y5n3f/?view_only=904174e2fc72413c9ecbe7b35a5507a1). To do this, we computed the weighted average of two effect-size distributions, one modeling the participant's prior belief (previously described), and one modeling the replication evidence. The replication evidence was modeled using the effect size (the distribution mean) and standard error from the replication results. Computing the weighted average of these two distributions (weighted by the ratio of the distributions' precision [the inverse of variance]) gave me a third distribution combining information from the participant's Phase I effect size estimates and the replication evidence (see Figure 2 for hypothetical example). The mean of this distribution indicated the calculated population effect size estimate, and the standard deviation was used to determine the calculated probability that the effect is greater than $d > .1$—the *Bayesian Posterior*. This *Bayesian Posterior* provides one standard against which to compare participants' *Actual Posteriors* (participants updated stated probability estimate that population effect is $d > .1$ after assessing the replication evidence.

Distinguishing Between Cases When Participants Should Adjust Upward Versus Downward

Several of my hypotheses make a distinction between when participants should, according to our Bayesian model, adjust their confidence (that an effect is greater than Cohen's *d*

Figure 2. Hypothetical Participant's Bayesian Posterior Based on their Actual Prior and the Replication Evidence



A hypothetical participant's prior belief of the effect size distribution (*green line*) is defined by a mean of .25 and an estimate that there is an 85% probability the (population) effect size is greater than $d = .1$, both reported by the hypothetical participant. Assuming a normal distribution, the standard deviation of this distribution is calculated to be .145. The replication evidence (*blue line*) is defined by a mean of .05 and a standard error of .05, both reported in the replication study report. These two distributions are used to compute the hypothetical participant's Bayesian posterior distribution (*purple line*). For this hypothetical participant, 27.2% of their Bayesian posterior distribution would be greater than $d = .1$, and therefore, their *Bayesian posterior* would be 27.2% (the area shaded in purple).

= .1) upward vs. downward. If a participant's *Bayesian posterior* was larger (smaller) than their *actual prior* for a given effect, I labeled this as a case when the participant "should" adjust upward (downward). (Throughout this manuscript, I use the phrase "*should* adjust upward/downward" in lieu of the more precise "would have adjusted upward/downward if acting

as a perfect Bayesian agent.") Therefore, the direction participants should adjust was not defined only by comparing the effect size from the original and replication studies, but also by their priors. There are many cases in which participants responding to evidence from the same study should adjust their confidence in opposite directions. Nevertheless, there was a direction the overwhelming majority of participants should have adjusted for most studies. For 15 of the 25 studies participants generally should have adjusted downward (i.e., the replication result effect size was lower than the center/mean of participants' priors), for 14 studies participants generally should have adjusted upwards, and for one study an approximately equal number of participants should have adjusted in each direction. Overall, participants provided 1,285 priors from which they should have adjusted downward, and 898 priors from which they should have adjusted upwards.

In some cases, the effect sizes and *actual priors* provided by participants made it impossible for me to calculate a *Bayesian posterior* (see Calculating Bayesian Posteriors section; these cases were not included in analyses for Hypothesis 2). In these cases, if the proportion of the replication evidence distribution that was above $d = .1$ was larger [smaller] than their *actual prior* for a given effect, this was classified as a case when the participants should adjust upward (downward). When both of these two methods could be used, they were in agreement about adjustment direction in 99.6% of cases.

Calculating *Predicted Posteriors*

For each study, participants in the prediction condition were asked to forecast the posterior they think they would provide for 6 possible outcomes (Cohen's *d* of -.15, .05, .20, .60, 1.20, and 1.80). To generate a single *predicted posterior,* I conducted a linear interpolation between the two posteriors provided for the possible outcomes that flanked the actual outcome

(see syntax here: https://osf.io/y5n3f/?view_only=904174e2fc72413c9ecbe7b35a5507a1). For example, if a replication produced a *d* of .10 (a third of the way between possible outcomes *d* = .05 and *d* = .20), we used the value that would fall a third of the distance along a linear trend between the posterior provided for *d* = .05 and that provided for *d* = .20.

To calculate participants' Predicted Posterior for each effect, I identified their predicted posteriors in response to the two hypothetical results scenarios that were the closest to the actual results. For example, when calculating a Predicted Posterior for a replication effect of *d* = .5, I would identify the participants prediction for hypothetical results of *d* = .2 and *d* = .6. If these predictions were, for instance, 40% and 60% I would calculate the Predicted Posterior for d = .5 as 55% because this is 40 plus 75% of the difference between 40 and 60 just as .5 is .2 plus 75% of the difference between .2 and .6.

Hypotheses

Hypotheses 3-5 are only tested for participants in the prediction condition because participants in the control condition did not provides ratings for all the variables in involves in these hypothesized relationships.

1) When participants should become more confident in effects their *Actual Posteriors* will be higher than their *Priors*. When participants should become less confident in effects their *Posteriors* will be lower than their *Priors*. In other words, participants will adjust their beliefs in the direction of the replication evidence.
2) When participants should become more confident in effects, their *Posteriors* will be lower than their *Bayesian Posteriors*. When participants should become less confident in effects, their *Posteriors* will be higher than their *Bayesian Posteriors*. In other words, participants will not sufficiently adjust their beliefs in either direction.

3) When participants should become more confident in effects, their *Posteriors* will be lower than their *Predicted Posteriors*. When participants should become less confident in effects, their *Posteriors* will be higher than their *Predicted Posteriors*. In other words, when results are merely hypothetical participants will predict that they will adjust their belief more than they actually do in response to real evidence.
4) When participants should become less confident in effects, higher *priors* will predict a greater *decrease* in ratings of *Replication Quality* after viewing replication evidence. In other words, participants who are more confident in original effects will, after the replication results suggest they should become less confident in the effect, become more critical of the replication studies. I do not have a prediction for this relationship when participants should become more confident in effects.
5) When participants should become less confident in effects, higher *priors* will predict a greater *increase* in ratings of *Replication Contextual Differences* after viewing replication evidence. In other words, participants who are more confident in original effects will, after the replication results suggest they should become less confident in the effect, identify more contextual differences between the original and replication studies. I do not have a prediction for this relationship when participants should become more confident in effects.
6) When participants should become less confident in effects, greater personal investment in the original finding being supported will predict a smaller magnitude decrease from *Prior* to *Posterior*. In other words, belief updating will be moderated by personal investment. I do not have a prediction for this relationship when participants should become more confident in effects.

7) Greater intellectual humility will predict greater belief updating in the direction participants should adjust their beliefs. In other words, belief updating (in the direction of the replication evidence) will be moderated by *Intellectual Humility*. While cases when participants should become less confident versus more confident in the effect were included in the same analysis, belief updating was calculated as *Posteriors* minus *Priors* when participants should have become more confident in effects and as *Priors* minus *Posteriors* when participants should have become less confident in effects. Therefore, positive values for belief updating indicate updating beliefs in the correct direction whereas negative value indicate belief updating in the incorrect direction.
8) Participants perceiving themselves to have greater *Bias Expertise* (average for ratings of professional expertise with confirmation bias and motivated reasoning) will predict greater belief updating in the direction participants should adjust their beliefs. In other words, belief updating (in the direction of the replication evidence) will be moderated by perceived *Bias Expertise*. While cases when participants should become less confident versus more confident in the effect were included in the same analysis, belief updating was calculated as *Posteriors* minus *Priors* when participants should have become more confident in effects and as *Priors* minus *Posteriors* when participants should have become less confident in effects. Therefore, positive values for belief updating indicate updating beliefs in the correct direction whereas negative value indicate belief updating in the incorrect direction.

Multilevel Modeling

All hypotheses were tested using multilevel modeling. Specifically, my data structure indicated using a cross-classified random effects model as responses were nested in stimuli (the

effects that participants evaluated) and participant but neither stimuli nor participant were nested in one another. Therefore, the multilevel model could incorporate two random effects (stimuli and participant) but only had two levels of data unlike hierarchical linear modeling which would have three levels for two random variables.

For all hypotheses, the initial model I fit to the data included the random effect of participant on intercept (e.g., the intercept of probability Cohen's d >.10) and the random effect of stimuli on both intercept and slope (e.g., the slope of *Prior* vs. *Posterior*). I selected this model on the basis of intuitive logic for how I expected responses to vary. I expected Stimuli to have an effect on intercept (because some effects would be generally more believable than others) and Stimuli would likely have an effect on slope (because studies would vary in their tendency to elicit entrenched beliefs that are less affected by replication evidence). However, incorporating the random effect of Stimuli on slope proved too complex for our sample size, providing minimal error reductions while dramatically shrinking degrees of freedom associated with the error term. Therefore, my MLM models only incorporate the random effects of Participant and Stimuli on the intercept.

The basic model can be expressed with classification notation (Browne et al. 2001) as

$$y_i = \beta_0 + \beta_1 x_i + u^{(2)}_{\text{participant}(i)} + u^{(3)}_{\text{stimuli}(i)} + e_i$$

$$u^{(2)}_{\text{participant}(i)} \sim \text{N}(0, \sigma^2_{u(2)})$$

$$u^{(3)}_{\text{stimuli}(i)} \sim \text{N}(0, \sigma^2_{u(3)})$$

$$e_i \sim \text{N}(0, \sigma^2_e)$$

where y_i denotes the outcome variable for the ith response, β_0 is the model intercept, x_i denotes the value of the predicter for the ith response, β_1 is the associated slope coefficient, $^{(2)}_{\text{participant}(i)}$ and $u^{(3)}_{\text{stimuli}(i)}$ denote the random effect of participant and stimuli for the ith response.

RESULTS

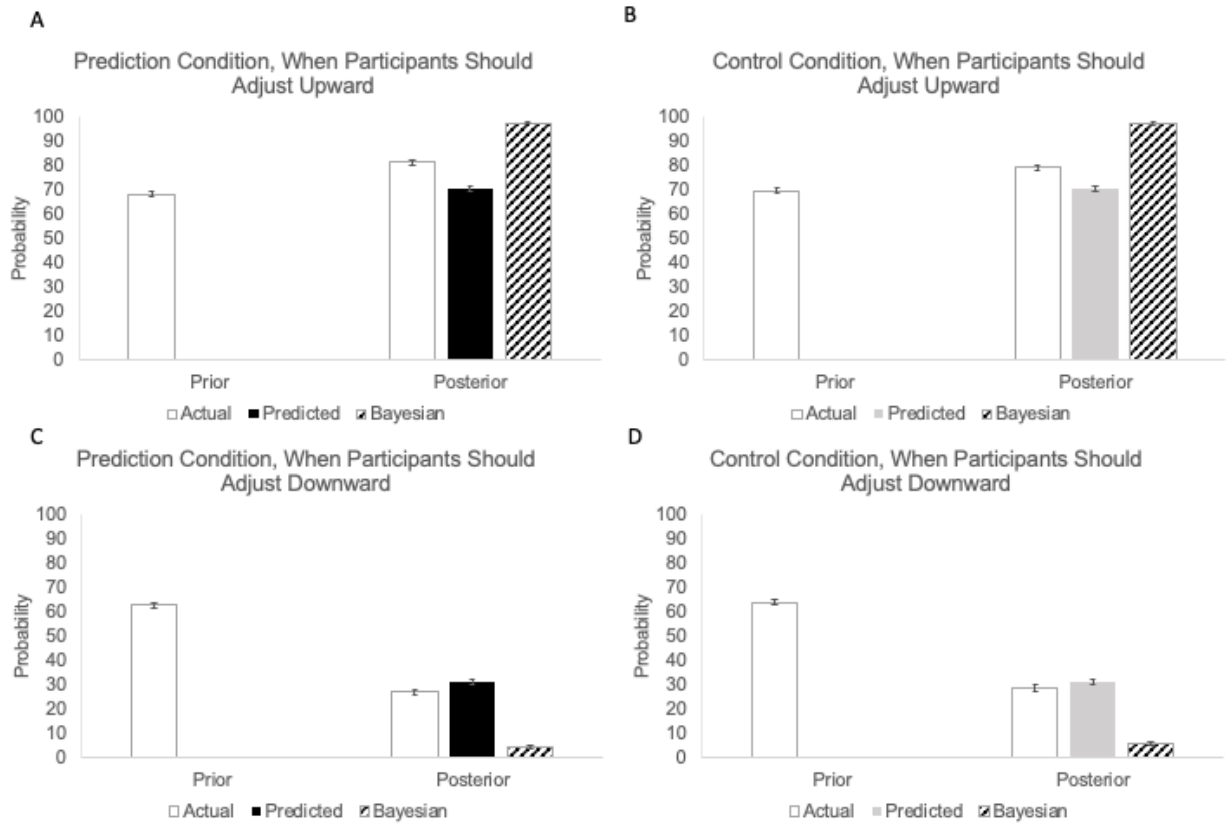
Hypothesis 1

The results were consistent with Hypothesis 1. Participants adjusted their estimated probability that the effect was real (i.e., $d > .1$) in the direction indicated by our Bayesian model. When the replication results indicated that participants should adjust downward, participants' *actual posteriors* ($M = 28.06$, $SD = 24.73$) were lower than their *actual priors* ($M = 63.16$, $SD = 24.37$), $b = -35.13$, $\beta = -.582$, 95% CI [-.609, -.555], $t(1851.8) = -42.77$, $p < .001$. When the replication results indicated that participants should adjust upward, participants' *actual posteriors* ($M = 79.98$, $SD = 21.15$) were higher than their *actual priors* ($M = 68.02$, $SD = 23.45$), $b = 11.98$, $\beta = .259$, 95% CI [.224, .293], $t(1123.41) = 14.66$, $p < .001$, (Figure 3).

Hypothesis 2

The results were consistent with Hypothesis 2. Participants did not update their estimates as much as our Bayesian model would indicate. When the replication results indicated that participants should adjust downward, participants' *actual posteriors* ($M = 28.06$, $SD = 24.73$), were not as low as their *Bayesian posteriors* ($M = 5.16$, $SD = 17.18$), $b = 23.80$, $\beta = .482$, 95% CI [.455, .509], $t(1676.6) = 34.40$, $p < .001$. When the replication results indicated that participants should adjust upward, participants' *actual posteriors* ($M = 79.98$, $SD = 21.15$) were not as high as their *Bayesian posteriors* ($M = 97.46$, $SD = 8.04$), $b = -15.84$, $\beta = -.424$, 95% CI [-.459, -.389], $t(1041.7) = -23.70$, $p > .001$. In an unpreregistered analysis, I also assessed this effect for the subset of cases

Figure 3. Comparison of Means for Hypotheses 1, 2, and 3



Note. Panel A displays the mean probability ratings for the prediction condition when participants should adjust their ratings upward. White bars represent participants' actual priors and posteriors, the black bar represents participants' predicted posteriors, and the striped bar represents participants' Bayesian posteriors. Panel B displays the same information as Panel A but for participants in the control condition. Here, the grey bar represents the predicted posterior provided by participants in the prediction condition, as participants in the control condition did not provide predictions. Panels C and D display the same information as A and B, respectively, but for cases when participants should adjust their ratings downward. Error bars indicate standard errors.

when participants gave the most positive possible evaluations (i.e., provided ratings at the most extreme point on our Likert-scale) of the replication study methodology during Phase II ("high quality" replication studies, and "minor contextual differences"). I conducted this analysis to examine whether the participants for whom replication results should have most strongly influenced belief updating—because their responses indicate the most confidence that the

replication was an appropriate test of the original effect—updated their beliefs as strongly as they should according to our model. Even when I limited our analyses to these 642 [629] cases in which participants should adjust downward [upward], the results still support Hypothesis 2, that participants do not update as much as our Bayesian model states they should. When the replication results indicated that participants should adjust downward, participants' *actual posteriors* (M = 16.72, SD = 20.67), were still not as low as their *Bayesian posteriors* (M = 3.96, SD = 15.88), b = 13.64, β = .340, 95% CI [.289, .391], $t(389.9) = 13.21$, $p < .001$. When participants should have adjusted upward, their *actual posteriors* (M = 83.16, SD = 19.57) were still not as high as their *Bayesian posteriors* (M = 98.06, SD = 7.82), b = -13.01, β = -.377, 95% CI [-.432, -.323], $t(613.0) = -13.64$, $p < .001$.

Hypothesis 3

Hypothesis 3 was not supported. Contrary to our hypothesis, participants updated *more* than they predicted they would. I tested this hypothesis in the prediction condition only, as these were the only participants that provided information to calculate their *predicted posteriors*. When the replication results indicated that participants should adjust downward, participants' *actual posteriors* (M = 28.06, SD = 24.73) were even *lower* than their *predicted posteriors* (M = 30.84, SD = 22.26), b = -3.02, β = -.060, 95% CI [-.096, -.024,], $t(1239.4) = -3.24$, $p = .001$. When the replication results indicated that participants should adjust upward, participants' *actual posteriors* (M = 79.98, SD = 21.15) were even *higher* than their *predicted posteriors* (M = 70.01, SD = 24.47), b = 10.69, β = .223, 95% CI [.184, .263], $t(898.8) = 11.04$, $p < .001$.

Hypothesis 4

The results largely failed to support my hypothesis. I observed little evidence that higher *actual priors* predicted a decrease in ratings of replication quality from Phase I to Phase II. I tested this hypothesis in the prediction condition only, as control participants did not provide ratings of replication quality in Phase I. To quantify changes in ratings of replication quality, we computed a difference score by subtracting participants' Phase I ratings from their Phase II ratings.

I ran an MLM model predicting changes in replication quality ratings from Phase I to Phase II from participants' *actual priors*. When the replication result indicated that participants should adjust downward, higher priors predicted a smaller increase in perceived replication quality, b = -.003, β = -.094, 95% CI [-.181, -.006], $t(586.7) = -2.10$, $p = .036$. That is, for participants who should have adjusted downward, those who expressed stronger prior beliefs in an effect were somewhat less likely to increase their ratings of the perceived quality of the replication in Phase II. This result provided mild support for my hypothesis. However, in contrast, a separate non-MLM analysis (averaging across participants' responses to individual studies), failed to support our hypothesis, β = -0.051, $p = .308$.

Because the previous MLM analysis obscures whether higher *actual priors* are associated with higher Phase I ratings or lower Phase II ratings, we ran an unpreregistered analysis assessing whether *actual priors* predicted Phase II ratings of replication quality controlling for Phase I ratings. My MLM model included Phase I replication quality, *actual priors,* and the interaction of these terms as fixed variables with Phase II replication quality as the outcome variable. The simple effect of *actual prior* was not significant b = -.002, β = -.068, 95% CI [-.146, .010], $t(598.0) = -1.70$, $p = .089$, nor was the interaction (β = .021, $p = .195$). Thus,

psychologists with similar Phase I ratings of replication quality did not have significantly different Phase II ratings depending on their *actual priors*. Despite the initial MLM analysis' weak support for our hypothesis, overall, these results do not provide more than trivial support (at best) for our hypothesis. It's also noteworthy that while we do not consider these results to be meaningful support for our hypothesis, these results also do not provide evidence contradicting a small effect in the direction of our hypothesis—the 95% confidence interval for my last analysis (the analysis I consider superior) includes small but meaningful effect sizes in the same direction as our hypothesis, lower bound of CI for $\beta = -.146$.

Hypothesis 5

The results did not support Hypothesis 5. I did not find evidence that higher priors predicted changes in perceptions of context differences. I tested this hypothesis in the prediction condition only, as control participants did not provide ratings of context differences in Phase I. To quantify changes in perceptions of context differences, I computed a difference score by subtracting participants' Phase I ratings of context differences from their Phase II ratings of context differences. I ran an MLM model predicting the difference score from participants' *actual priors*. When the replication result indicated that participants should adjust downward, *actual priors* were not associated with changes in perceived contextual differences**,** $\beta = -.044$, $p = .302$.

Hypothesis 6

Results did not support Hypothesis 6. When replication results suggested participants should adjust their beliefs downward, people who were more personally invested in a finding adjusted *more*, not less. I ran an MLM model predicting adjustment (*actual posterior – actual prior*) from personal investment. When the replication result indicated that participants should

adjust downward, more personal investment predicted *greater* downward adjustment, b = -8.42, β = -.155, 95% CI [-.207, -.103], t(1,202.1) = -5.85, p < .001. Contrary to my hypothesis, this finding suggests that psychologists who are more personally invested in a finding show a greater decrease from their priors to posteriors when replication results warrant such a change.

The previous analysis obscures whether higher personal investment is associated with higher *actual priors* or lower *actual posteriors* (the latter being more relevant to our hypothesis). To more explicitly dissect the relationship between personal investment and belief updating, I ran an unpreregistered analysis assessing whether personal investment predicted *actual posteriors* when controlling for *actual priors*. My MLM model included *actual prior* and personal investment, and the interaction of these terms as fixed variables with *actual posterior* as the outcome variable. The simple effect of personal investment, for the average *actual prior*, was significant, b = 2.20, β = .052, 95% CI [.002, .102], t(1,104.1) = 2.02, p = .043, while the interaction of personal investment and *actual prior* was not significant, β = -.035, 95% CI [-.082, .011], p = .138. When participants should adjust downward, these results suggest that for psychologists with equal priors, one should predict slightly higher *actual posteriors* for psychologists with more personal investment in an effect. This result suggests a caveat to the original finding that participants with more personal investment show greater belief updating; participants who are equally credulous before evaluating replication evidence, show slightly *less* belief updating when they are more personally invested in the effect.

Hypothesis 7

Results did not support Hypothesis 7. Expertise with bias was unrelated to the degree to which participants updated their beliefs. My outcome variable was computed as *actual posterior – actual prior*, in cases when participants should have adjusted upward, and as *actual prior –*

actual posterior in cases when participants should have adjusted downward. I ran an MLM model predicting the extent to which participants adjusted in the correct direction from expertise. The association was not significant, β = .009, 95% CI [-.034, .052], p = .680, indicating that expertise regarding bias was not predictive of belief updating. Furthermore, the 95% confidence interval suggests that if any relationship exists between perceived expertise with bias and belief updating, the effect size is likely trivial.

Hypothesis 8

The results supported Hypothesis 8. Participants with greater intellectual humility showed slightly greater belief updating in the correct direction. My outcome variable was computed as *actual posterior – actual prior*, in cases when participants should have adjusted upward, and as *actual prior – actual posterior* in cases when participants should have adjusted downward. I ran an MLM model with intellectual humility as the fixed variable predicting the extent to which participants adjusted in the correct direction. As anticipated, greater intellectual humility predicted greater belief updating in the correct direction, b = 5.49, β = .086, 95% CI [.043, .128], t(978.8) = 3.93, p < .001.

I followed-up on this finding with an unpreregistered analysis assessing whether intellectual humility predicted *actual posteriors* when participants started with similar priors (i.e., controlling for the effect of *actual prior* on *actual posterior*). My MLM model included *actual prior*, intellectual humility, and the interaction of these terms as fixed variables with *actual posterior* as the outcome variable. When the replication result indicated that participants should adjust downward, the simple effect of intellectual humility was significant, b = -6.61, β = -.121, 95% CI [-.211, -.050], t(411.8) = -3.19, p = .002. When the replication result indicated that participants should adjust upward, the

simple effect of intellectual humility was also significant, $b = 3.59$, $\beta = .085$, 95% CI [.013, .157], $t(306.1) = -2.33$, $p = .021$. Given the nonsignificant interactions (regardless of the normative adjustment direction) of *actual prior* and intellectual humility ($\beta = .034$, $p = .359$; $\beta = .019$, $p = .580$), these simple effects suggest that for psychologists with equal priors, one should predict slightly greater belief updating from psychologists with greater intellectual humility.

DISCUSSION

For replication studies to be an effective component of the scientific method and a tool for scientific self-correction, scientists must update their beliefs in the direction consistent with evidence from replication studies. In the present study, psychologists did update their beliefs in the appropriate direction—regardless of if replication results suggested downward or upward adjustments in confidence—and the magnitude of these adjustments were rather substantial. However, according to my Bayesian model, psychologists did not update their beliefs as strongly as was warranted by the evidence. This finding held even when the analyses only included participants whose assessments of the replication studies were the most positive assessments allowable by our scales ("high quality" replication study and "minor contextual differences" between the original and replication study). Therefore, the results suggest there is some inefficiency in on of the mechanisms for scientific self-correction and this can contribute to unwarranted skepticism and unwarranted credulity regarding many psychological effects.

Overall, I did not observe evidence that motivated reasoning is a mechanism contributing underadjusment in belief updating. Participants in the prediction condition did not update their beliefs less than they predicted they would (for downward or upward adjustments in confidence), participants with strong priors in support of effects were not more likely to become critical of replication study methodology when the results conflicted with their preexisting beliefs, and participants who were more personally invested in findings gaining future support were not more

resistant to changing their beliefs than participants who were personally invested in findings being refuted. The latter two null effects persisted even when controlling for participants priors.

With regards to personal differences moderating belief updating, there was not any evidence that bias expertise moderated belief updating but I did observe evidence that greater intellectual humility predicts slightly more belief updating.

Limitations

Our sample of psychologists makes this study particularly susceptible to demand characteristics. When evaluating replications of studies previously evaluated by participants, it is quite possible that many participants correctly guessed that the study was investigating belief updating. If such a realization made participants conscientious of avoiding belief preservation behaviors, it could be that participants compensated for such a concern (possibly due to a desire to present as less biased) by adjusting their beliefs more than they might normally adjust their beliefs.

Conversely, aside from demand characteristics, it is possible that psychologists adjusted their beliefs less than they would in their natural environment because they read summaries of replication studies rather than full replications studies. For example, perhaps not being able to fully assess the rationale for the hypotheses of original studies influence participants to update their beliefs less they would if they could better identify reasons to be critical of the theoretical justification for original effects.

Another limitation of this study is that our Bayesian model makes a couple of assumptions that may not be fully justified. Most critically, our Bayesian model assumes participants view replication studies as a direct test of the original effect. This assumption is unlikely to be strictly true for any replication study and our finding that participants did not update their beliefs as

much as they "should" have (according to our model and only considering accuracy motivations) may be due, to some extent, due to participants rationally devaluing the weight of replication evidence because these studies were not viewed to be testing the *exact* same effect as the original finding. In other words, participants may have attributed part of the disparity between original results and replication results to a lack of support auxiliary hypotheses (e.g., the assumption that the replication study does not have hidden moderators) rather than a lack of support for the central hypothesis that the original finding was accurate. We attempted to address the possibility that participants may have updated their beliefs less because of their critiques of replications' methodology. Participants who had the most positive attitudes regarding the replication studies' ability to test the original finding did adjust more than those who did not have these attitudes but they still substantially underadjusted relative to adjustment prescribed by my Bayesian model. While participants indicating that replication studies are high quality and have minor contextual difference from the original study is by no means an indication the replication studies are perfect, these findings still suggestive that underadjusment in belief change cannot be fully attributed to perceiving the replication studies to not be an appropriate test of the original effect. Despite the fact that our study assessed an important auxiliary hypothesis that could have been perceived to account for original effect versus replication effect disparities, there are other perceptions of the replication study that I did not assess that could account for some of the observed insufficient belief updating.

My Bayesian model also assumed a normal distribution for participants' prior distributions for confidence in the original effects. While a normal distribution seems like a sounder assumption than alternatives (e.g., leptokurtic distributions) it is certain that

participants mental constructions of their prior distributions do not assume a perfectly normal shape. Despite this concern, while other assumptions for prior distribution shapes would result in result in our model describing participants as underadjusting to a smaller degree, prior distributions would need to have an SD approximately 2.75 times smaller than those assumed by our model for belief updating to be comparable to that prescribed by a model. While different model assumptions could reasonably deflate our effect size estimate for Hypothesis 2, assumptions that would eliminate this effect seem unfounded.

Conclusion

For psychology to be self-correcting (among other conditions): 1) replications must be regularly conducted, 2) psychologists must be exposed to the results of replications studies, and 3) psychologists must be persuaded to update their beliefs as warranted by replication evidence. This research suggests this third condition is meet to an extent that could—contingent on the extent to which the first two conditions are met—allow for meaningful (though suboptimal) scientific self-correction. However, our finding that psychologists would be justified in engaging in substantially stronger belief updating suggests a critical barrier that could be severely impeding the efficiency self-correction in psychology. When psychologists do not completely abandon confidence in effects for which 7,000 participant replication studies—studies that the psychologists in question rate as high-quality tests of the original effects—find the effect to be negligible or nonexistent, this raises concern that such effects will continue be faulty building blocks for future research, and taught to the general public, long after it has ceased to be prudent to consider these effects credible.

REFERENCES

Ballarini, C., & Sloman, S. A. (2017). Reasons and the "Motivated Numeracy Effect.". *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, 1580-1585.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling*, *1*(2), 103-124.

Chambers, C. D., Feredoes, E., Muthukumaraswamy, S. D., & Etchells, P. (2014). Instead of" playing the game" it is time to change the rules: Registered Reports at AIMS Neuroscience and beyond. *AIMS Neuroscience*, *1*(1), 4-17.

Djulbegovic, B., & Hozo, I. (2007). When should potentially false research findings be considered acceptable? *PLOS Medicine*, *4*(2), e26.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891-904.

Frank, M. C., & Saxe, R. (2012). Teaching replication. *Perspectives on Psychological Science*, *7*(6), 600-604.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), e124.

Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, *1*(1), 54-86.

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior & Human Decision Processes*, *56*, 28.

Koole, S. L., & Lakens, D. (2012). Rewarding replications a sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*(6), 608-614.

Kruglanski, A. W., & Webster, D. M. (1996). Motivated Closing of the Mind:" Seizing" and" Freezing. *Psychological Review*, *103*(2), 263-283.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480.

Leary, M. R., Diebels, K. J., Davisson, E. K., Isherwood, J. C., Jongman-Sereno, K. P., Raimi, K. T., ... Hoyle, R. H. (2016). Cognitive and interpersonal features of intellectual humility. Manuscript under review Durham, NC: Duke University.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research*, *1*(2), 161-175.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.

Panagiotou, O. A., & Ioannidis, J. P. (2012). Primary study authors of significant studies are more likely to believe that a strong association exists in a heterogeneous meta-analysis compared with methodologists. *Journal of Clinical Epidemiology*, *65*(7), 740-747.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *PNAS Proceedings of the National Academy of Sciences of the United States of America*, *113*(23), 6454–6459.

APPENDICIES

Appendix A

Original Study

Citation: Huang, Y., Tse, C. S., & Cho, K. W. (2014). Living in the north is not necessarily favorable: Different metaphorical associations between cardinal direction and valence in Hong Kong and the United States. European Journal of Social Psychology, 40, 360-369.
Summary: People in the United States and Hong Kong have different demographic knowledge that may shape their metaphoric association between valence and cardinal direction (North/South). Participants were presented with a blank map of a fictional city and were randomly assigned to indicate on the map where either a high-SES or low-SES person might live.
Sample: 180 participants in the United States and Hong Kong.

Key Finding: Participants from the US (compared to participants in HK) were more likely to think the high SES person lived further North than the low SES person. There was an interaction between location (US vs. HK) and SES (high vs. low) in predicting how far North participants expected the person would live, $F(1, 176) = 20.39$, $p < .001$, $\eta_p^2 = 0.10$, $d = .68$, 95% CI [.38, .98] .

US participants expected the high-SES person to live further north (M = +0.98, SD = 1.85) than the low-SES person (M = -.69, SD = 2.19). Conversely, HK participants expected the low-SES person to live further north (M = +0.63, SD = 2.75) than the high-SES person (M = -0.92, SD = 2.47).

Replication

Key Finding from Original Study: Participants from the US (compared to participants in HK) were more likely to think the high SES person lived further North than the low SES person.

Sample: Data will be collected from approximately 47 unique samples. The overall N is expected to be approximately 2,350.
Known Differences from Original: Original participants were asked to guess the purpose of the study afterward but none did, and we will not be including that item.

The original was presented on pencil-and-paper, and participants drew an "X" on the map. However, the replication will be on a computer and participants will click to indicate the location on the map. With a monitor presentation, this also means the study will be completed on a vertical display as opposed to a horizontal monitor. The original authors suggest that this may be particularly important because associations between "up" and "good" or "down" and "bad" may interfere with any North/South associations. As such, at eight data collection sites, participants will be randomly assigned to complete the slate on a regular monitor or on a Microsoft Surface tablet that is resting on the table, like a paper-pencil administration.

Lastly, the original analysis emphasized t-tests against zero, whereas the replication will focus specifically on the difference between conditions (e.g., independent samples t-test).

Appendix B

THE UNIVERSITY OF ALABAMA | *Office of the Vice President for* Research & Economic Development
Office for Research Compliance

August 13, 2018

Alexa Tullett, Ph.D.
Assistant Professor
Department of Psychology
College of Arts & Sciences
The University of Alabama
Box 870348

Re: IRB # 16-OR-273-ME-R2 "Perceptions of Psychology Findings"

Dear Dr. Tullett:

The University of Alabama Institutional Review Board has granted approval for your renewal application. Your renewal application has been given expedited approval according to 45 CFR part 46. You have also been granted the requested waiver of informed consent. Approval has been given under expedited review category 7 as outlined below:

(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

Your application will expire on August 12, 2019. If your research will continue beyond this date, complete the relevant portions of the IRB Renewal Application. If you wish to modify the application, complete the Modification of an Approved Protocol Form. Changes in this study cannot be initiated without IRB approval, except when necessary to eliminate apparent immediate hazards to participants. When the study closes, complete the appropriate portions of the IRB Study Closure Form.

Should you need to submit any further correspondence regarding this proposal, please include the above application number.

Good luck with your research.

Sincerely,

358 Rose Administration Building | Box 870127 | Tuscaloosa, AL 35487-0127
205-348-8461 | Fax 205-348-7189 | Toll Free 1-877-820-3066