Revisiting the Folk Concept of Intentionality: Replications of Malle and Knobe (1997)

*Subramanya Prasad Chandrashekar
Department of Psychology,
Norwegian University of Science and Technology (NTNU)
prasad.chandrashekar@gmail.com / prasad.chandrashekar@ntnu.no
ORCID: 0000-0002-8599-9241

*Yee Yan Chan (cyy.nice@gmail.com / u3532660@connect.hku.hk),
*Kar Lam Cheng (cckl3518@connect.hku.hk)
Department of Psychology, University of Hong Kong, Hong Kong SAR

*Donna Yao
jyaoab@connect.ust.hk

** Chung Yan Sharon Lo (cylosharon@gmail.com / u3517887@connect.hku.hk),
** Tsz Chun Anson Cheung (tszchun.anson@icloud.com / tcanson@connect.hku.hk),
**Hoi Yan Sandy Tang (hoiyan9@hotmail.com / u3505661@connect.hku.hk),
**Yuen Ting Andy Leung (andy.ytleung@gmail.com / andii@connect.hku.hk),
**Chi N Tsoi (neotsoi@gmail.com / chintsoi@connect.hku.hk)
Department of Psychology, University of Hong Kong, Hong Kong SAR

Bo Ley Cheng (boleystudies@gmail.com), Ka Wai Ng (kawaing@hku.hk)
Department of Psychology, University of Hong Kong, Hong Kong SAR

^Gilad Feldman
Department of Psychology, University of Hong Kong, Hong Kong SAR
gfeldman@hku.hk
ORCID: 0000-0003-2812-6599

In press at Journal of Experimental Social Psychology

Accepted for publication on June 2, 2022

*Contributed equally, joint first author
** Contributed equally, joint fifth author
^Corresponding author
Word: abstract – 148, manuscript - 3718

Author Bios:

Subramanya Prasad Chandrashekar is a postdoctoral fellow with the Department of Psychology at the Norwegian University of Science and Technology (NTNU). His research focuses on social status, lay-beliefs, and judgment and decision-making.

Donna Yao was a research fellow and a graduate of the Hong Kong University of Science and Technology management department in the year 2019.

Chan Yee Yan, Cheng Kar Lam, Chung Yan Lo, Tsz Chun Cheung, Hoi Yan Tang, Yuen Ting Leung, and Chi N Tsoi were undergraduate students at the University of Hong Kong in the year 2018.

Bo Ley Cheng and Ka Wai Ng were teaching assistants at the University of Hong Kong psychology department in the year 2018.

Gilad Feldman is an assistant professor with the University of Hong Kong psychology department. His research focuses on judgment decision-making.

Declaration of Conflict of Interest:

The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial disclosure/funding:
The author(s) received no financial support for the research and/or authorship of this article.

Acknowledgments or Author Notes:

Subramanya Prasad Chandrashekar would like to thank Institute of International Business and Governance (IIBG), established with the substantial support of a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/IDS 16/17), for its support.
We would like to thank Adrien Fillon and Ignazio Ziano for their helpful feedback on drafts of this manuscript.

Corresponding author

Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR
gfeldman@hku.hk

Authorship declaration

A team comprising of Yee Yan Chan and Kar Lam Cheng designed and wrote the pre-registrations for Study 2. A team comprising of Chung Yan Sharon Lo, Tsz Chun Anson Cheung, Hoi Yan Sandy Tang, Yuen Ting Andy Leung, and Chi N Tsoi designed and wrote the pre-registrations for Study 1. Bo Ley Cheng, Ka Wai Ng, and Gilad Feldman guided the two teams. Gilad Feldman conducted the pre-registration, and ran data collection. The two teams analyzed and wrote initial reports of the findings.
Donna Yao did a first integration of the two projects to a manuscript. Prasad followed on the initial work to verify, add analyses, and complete the manuscript for submission. Given substantial crucial feedback received from journal and community peer-review, Prasad completely redid the analyses and rewrote the manuscript.
Following feedback and peer review, prasad finalized and Gilad edited the manuscript and revisions for submission.

Contributor Roles Taxonomy

In the table below, employ CRediT (Contributor Roles Taxonomy) to identify the contribution and roles played by the contributors in the current replication effort. Please refer to the url (https://www.casrai.org/credit.html) on details and definitions of each of the roles listed below.

Role	Subramanya Prasad Chandrashekar,	Gilad Feldman	Donna Yao	Chan Yee Yan, Cheng Kar Lam, Chung Yan Lo, Tsz Chun Cheung, Hoi Yan Tang, Yuen Ting Leung, and Chi N Tsoi	Bo Ley Cheng, Ka Wai Ng
Conceptualization		X			
Pre-registrations		X		X	
Data curation		X			
Formal analysis	X		X	X	
Funding acquisition		X			
Investigation				X	
Methodology				X	
Pre-registration peer review / verification				X	X
Data analysis peer review / verification	X		X	X	
Project administration		X			X
Resources		X			
Software				X	
Supervision		X			X
Validation	X	X			
Visualization	X				
Writing-original draft	X				
Writing-review and editing	X	X			

Abstract

Malle and Knobe's (1997) Study 1 found that people exhibit a shared understanding of intentionality and apply it consistently in their judgments. The study found that different people tend to judge intentionality similarly, and that intentionality ratings were consistent across a set of behaviors from an actor's or an observer's perspective. Additionally, the presence or absence of a definition of intentionality as part of the study instruction did not seem to affect the intentionality judgments. We conducted two pre-registered replications of Study 1 (N=46; N=817). The replication results provide support for the findings of the original study. Consistent with the original study, we found high inter-rater agreement across perspective (actor vs. observer) experimental conditions; and no evidence for differences in intentionality ratings depending on whether a definition of intentionality was provided. We observed that actor perspective led to a higher average rating of intentionality than the observer perspective. Materials, data, and code are available on: https://osf.io/4q5ce/

Keywords: folk concept, intentionality, attribution, pre-registered replication, actor-observer

Revisiting the Folk Concept of Intentionality:

Replication of Malle and Knobe (1997)

Compared with other species, humans are most adept at coordinating with other members of the species, and have evolved to inhabit societies with complex rules. Much of this adeptness comes from the ability of humans, with a fair degree of accuracy, to discern information about the social world around them (Tomasello et al., 2005). Understanding intentions of other individuals is fundamental to this ability (Searle, 1995).

Research on attribution aims to understand people's lay perceptions, tackling topics such as the understanding of causes and consequences of human behavior. In his seminal work, Heider (1958) identified intentionality as one of the key factors that people use when making attributions. Judgments of intentionality serve to form associations between the perceived mind and observed behavior. The perception of intention has implications for outcomes such as judgments of blame, helping (Swap, 1991), etc. Naïve people distinguish between intentional and unintentional behaviors (e.g., intentional murder vs. manslaughter) and assign harsher penalties to intentional misbehaviors. Ample research has been devoted to better understand judgments of intentionality.

Malle and Knobe's (1997) work established an important need to investigate the folk concept of intentionality. In their work, termed as the *folk concept of intentionality*, the authors found that people showed substantial agreement with intentionality judgments. They concluded

that an intentional action includes: 1) a desire on the part of the actor for an outcome; 2) the belief that the action leads to that outcome; 3) the intention to act; and 4) awareness of performing that action.

The chosen study for replication: Malle and Knobe (1997)

Malle and Knobe's (1997) work was the first to investigate intentionality as a shared folk concept. The paper has contributed to later works on intentionality in multiple disciplines, including cognitive psychology, law, philosophy, and communications (e.g., Dennison, 2007; Sloman et al., 2012; Vangelisti & Young, 2000; Zalla & Leboyer, 2011).

Before Malle and Knobe's (1997) paper, research work offered a conceptual analysis of intentional behavior with limited empirical investigations (Heider, 1958; Jones & Davis, 1965). Malle and Knobe's (1997) work was influential in addressing this limitation by empirically investigating the necessary and sufficient conditions to regard a behavior as intentional. Furthermore, the work has contributed to identifying the limitations of traditional attributional concepts for a folk explanation of behaviors (Malle, 1999). For instance, one such limitation is the view that people think of human behavior in simple person-versus-situation terms. Malle and Knobe's (1997) work has formed the basis for further theoretical assertions in the form of the *folk-conceptual theory of explanation*, which uses the sorting of naturally occurring verbal descriptions of behavior into identifiable theoretical categories (Malle, 2004; Malle et al., 2000). Additionally, Malle and Knobe's (1997) work has led to further works in moral psychology. For example, recent research on the folk understanding of intentionality, choice, and free will have

contributed to a better understanding of people's judgments of moral responsibility (Feldman et al., 2014; Guglielmo et al., 2009; Knobe, 2010; Malle et al., 2012).

Study 1 of Malle and Knobe (1997) tested the premise of the *folk concept of intentionality*: whether people show substantial agreement when judging intentionality on a variety of behaviors, or whether intentionality judgments vary. Following Study 1, the authors empirically attempted to elucidate the necessary and sufficient components of the folk concept of intentionality. We were especially interested in revisiting Study 1 as it tested a fundamental assumption of the proposed folk concept.

The article has had a major impact on social psychology and experimental philosophy research. Its paradigm has also been used to examine whether there are cultural differences in the concept of intentionality (Ames et al., 2001; Ohtsubo, 2007). At the time of writing (February 2022), Google Scholar reported 919 citations of the article, indicating that it is foundational to a broader understanding of how people explain behavior. To the best of our knowledge, there are no published direct replications of Malle and Knobe (1997). Heeding increasing calls to revisit and reassess published work (Gelman & Loken, 2013; Nosek & Lakens, 2014), we embarked on pre-registered replications of Study 1 in Malle and Knobe (1997).

Overview of replications

We contacted the authors to obtain the materials and data collected during the original study[1]. Table 2 summarizes the results of the original study. We conducted two pre-registered direct replications of the Malle and Knobe (1997) Study 1. Our study 1 recruited a small and specialized Hong Kong undergraduate students sample as a pre-test, and Study 2 involved a larger American sample recruited on Amazon's Mechanical Turk (MTurk) platform. We made adjustments to the preregistration data analysis plan (see Table S1 of the supplementary materials for details). Both samples were collected in the year 2018.

Open science, preregistrations, and disclosures

Data, code, and materials were made available on the Open Science Framework (project: https://osf.io/4q5ce/; preregistration Study 1: https://osf.io/wtpsr/; preregistration Study 2: https://osf.io/mt5jq/).

All studies, participants, measures, manipulations, and exclusions conducted for this investigation are reported, all inferential tests not explicitly marked "exploratory" were pre-registered with power analyses, and data collection was completed before hypothesis testing. All *t*-tests were two-tailed, and α was set at .05.

[1] We would like to thank the authors of the original study for providing the study materials and raw data.

Method

Power analysis

We used G*Power (Faul et al., 2007) to estimate the effect sizes of the perspective and definition conditions on intentionality rating in the original study. As Cohen's *d* is below .20 for small effects (Cohen, 1988), to obtain power = .80 and $\alpha = .05$, a sample size of 788 was required. We conducted Study 2 with a well-powered sample size. For Study 1, however, the sample size was determined by the availability of sample (see the details below), hence we considered Study 1 as a pre-test.

Participants

Study 1: Hong Kong undergraduate students

The first replication serving as a pre-test in a small and well-informed sample was conducted with undergraduate students who took a psychology course at a university in Hong Kong. Students were randomly assigned into groups of 3 to 6. Each group conducted a replication of previously published findings, and one of the replications was our Study 1. Students who were not assigned to design the replication for this study served as the participants, and they had no knowledge of the study design or hypotheses prior to participation. Those who were in charge of the replication also took part and were included in the dataset, but were later excluded from the data analysis. The final sample included 46 participants (15 males, 31 females; $M_{age} = 20.15$, $SD_{age} = 1.01$).

Study 2: US Americans on MTurk

A total of 835 US American participants were recruited on MTurk using TurkPrime/CloudResearch (Litman et al., 2017). The final sample included 817 participants[2] (334 males, 483 females; M_{age} = 39.6, SD_{age} = 12.6).

Design, procedure, and measures

Study 1

Study 1 was conducted as part of a combined set of experiments by different researchers presented in a randomized order. Its design and materials were almost identical to the original study, except for some minor modifications, as elaborated below. Participants were randomly assigned to one of four conditions in a between-subjects design: 2 (perspective: actor vs. observer) × 2 (with vs. without definition of intentionality). Participants were presented with the 20 behavioral statements used in the original study in a random order in all conditions.

Instructions to the participants in the actor-perspective condition read: "Please look at the 20 statements below. Each statement describes you doing something. Your task is to rate whether you would do that *intentionally*." The statements were described from the first-person perspective (e.g., "I won a prize in the lottery."). Participants in the observer-perspective condition were instructed: "Please look at 20 statements below. Each statement describes Anne doing something. Your task is to rate whether Anne does what she does *intentionally*." The

[2] A total of 18 participants were excluded based on our pre-registered exclusion criteria. Specifically, eight participants were excluded due to not completing the study, and 10 participants were excluded because they reported a low understanding of the English used in the study (i.e., lower than 5 (*fair*) on a 7-point Likert scale [1 = *very bad*, 7 = *very good*]).

statements were described from a third-person perspective (e.g., "Anne won a prize in the lottery. ") (see Table 1). All participants were asked to rate the intentionality on an 8-point Likert scale (0 = *Not at all*, 7 = *Completely*).

In addition, participants in the with-definition condition were presented with a working definition of intentionality before they rated the 20 behaviors. Participants in the actor perspective read: "What do we mean by *intentional*? This means that you had a reason to do what you did and that you chose to do so." Participants in the observer perspective read: "What do we mean by *intentional*? This means that the person had a reason to do what she did and that she chose to do so." Those in the without-definition condition were not presented with the definition.

Table 1. *Behavioral statements in Studies 1 and 2.*

	Actor Perspective	Observer Perspective
Item 1	I am sweating.	Anne is sweating.
Item 2	I was yawning during the lecture.	Anne was yawning during the lecture.
Item 3	I was grinding my teeth during the test.	Anne was grinding her teeth during the test.
Item 4	I had a craving for cherries after dinner.	Anne had a craving for cherries after dinner.
Item 5	I believed that I had the flu.	Anne believed that she had the flu.
Item 6	I won a prize in the lottery.	Anne won a prize in the lottery.
Item 7	I am in a great mood today.	Anne is in a great mood today.
Item 8	I am infatuated with someone.	Anne is infatuated with someone.
Item 9	I was worrying about the test results.	Anne was worrying about the test results.
Item 10	I got admitted to a university.	Anne got admitted to a university.
Item 11	I interrupted my mother.	Anne interrupted her mother.
Item 12	I ignored someone's arguments.	Anne ignored someone's arguments.
Item 13	I drove way above the speed limit.	Anne drove way above the speed limit.
Item 14	I applauded the musicians.	Anne applauded the musicians.
Item 15	I greeted my uncle politely.	Anne greeted her uncle politely.
Item 16	I refused the salesman's offer.	Anne refused the salesman's offer.
Item 17	I stole a pound of peaches.	Anne stole a pound of peaches.
Item 18	I asked someone out for dinner.	Anne asked someone out for dinner.
Item 19	I invited someone to have lunch with me.	Anne invited someone to have lunch with her.
Item 20	I watered my new plants.	Anne watered her new plants.

It should be noted that we made some minor modifications to the original statements. We replaced the target of actions in the original items (i.e., Ben, Mike, and Sue) with "someone" to try and avoid any issues related to gender and/or sexuality that perhaps were not considered a concern in the 1990s. We also changed Princeton in "Anne got admitted to Princeton" into "Anne got admitted to a university" due to our studies' changed location and context.

At the end of the study, participants reported their seriousness in completing the questionnaire and their English comprehension ability.

Study 2

Study 2 was identical to Study 1, except that we added a comprehension question between the manipulation of intentionality definition and the behavioral statements to adjust to online data collection. After reading the definition of intentionality, participants read, "To make sure you understood our definition, please answer the following comprehension question - What do we mean by intentional? This means that - ". Participants were asked to choose the correct answer out of three options: (1) You had a reason to do what you did, (2) You chose to do what you did, and (3) You had a reason to do what you did and that you chose to do so. A comprehension question helped ensure that participants correctly understood the definition and were doing the survey seriously. Participants had to answer correctly in order to proceed.

We characterize the current replication as a "very close replication" based on the framework for classification of replications using the criteria by LeBel et al. (2018) (see Table S15 and S16 in the supplementary material).

Analysis strategy

We conducted three sets of analyses. First, we tested the degree of agreement in judging twenty behaviors for their intentionality. The measures of the agreement include 1) Average inter-rater correlation, 2) Average rater whole correlation (i.e., average correlation of a person's rating with the remaining group), 3) Intraclass Correlation Coefficient (ICC), 4) Inter-rater reliability (coefficient alpha measure). We report the results based on the responses across the whole sample and within experimental conditions (Definition and Perspective). Second, we

calculated Lin's concordance correlation coefficient as a measure of agreement between actor-observer perspectives and between the presence or absence of definition available to the participants. Third, as part of an exploratory data analysis, we investigated actor-observer differences in the intentionality ratings.

The pre-registered analytic approach to test the actor-observer differences included 20 pairwise independent *t*-tests. However, we report results based on linear mixed-effects models (LMEM) to improve the shortcomings of the analyses proposed as part of the preregistration[3]. The results are based on the *lme4* package in R (Bates et al., 2015). The significance of fixed effects was assessed via Satterthwaite's degrees of freedom (Kuznetsova et al., 2017). The mixed-effects models adjusted for covariates at Level 2 (i.e., item number, sample identifier, and participants' ID were treated as random intercepts). The LMEMs allow for combined data analysis while accommodating the nested nature of the data samples (Judd et al., 2012).

Results

We summarized descriptive statistics of the measures across Study 1 and 2 in Table S5–S8 of the supplementary material.

[3] We proposed ANOVA as part of the preregistration. However, we report results based on linear mixed-effects models (LMEM) to improve on the shortcomings of the pre-registered analyses and to address feedback given in the peer-review process.

People's agreement on intentionality ratings

In Study 1, any two participants showed an average intercorrelation of $r = .40$, and any one person showed an average correlation of $r = .62$ with the remaining group. Additionally, we found very high inter-rater reliability (coefficient α measure) = 0.97 and ICC = 0.40 (see Table 2 for detailed results).

In Study 2, any two participants showed an average intercorrelation of $r = .59$, and any one person showed a high average correlation of $r = .77$ with the remaining group. As a measure of reliability, we found very high inter-rater reliability (Cronbach's α measure) = 0.99. Furthermore, we found ICC = 0.58. The measure did not vary between the actor perspective (.77) or observer perspective (.77) conditions, or between the conditions with definition (.79) or without definition (.75). All correlations were relatively high.

Table 2. Measures of agreement among participants' ratings

	Original study	Study 1	Study 2
Sample size	104	46	817
Average inter-rater correlation	0.64	0.40	0.59
Average rater–whole correlation	0.80	0.62	0.77
Intraclass Correlation Coefficient (ICC)	0.58	0.39	0.58
Inter-rater reliability (*coefficient alpha* measure*)*	0.99	0.97	0.99
By perspective			
Actor			
Average inter-rater correlation	0.60	0.38	0.59
Average rater–whole correlation	0.77	0.60	0.77
ICC	0.56	0.37	0.58
Observer			
Average inter-rater correlation	0.65	0.40	0.59
Average rater–whole correlation	0.80	0.62	0.77
ICC	0.60	0.39	0.58
By definition			
With definition			
Average inter-rater correlation	0.63	0.46	0.63
Average rater –whole correlation	0.79	0.66	0.79
ICC	0.59	0.45	0.61
Without definition			
Average inter-correlation	0.64	0.33	0.56
Average rater–whole correlation	0.79	0.56	0.75
ICC	0.57	0.33	0.55

Note: Average rater–whole correlation = Average correlation of a person's rating with remaining group.

Agreement across perspectives and definition

We calculated average ratings across 20 behaviors within each experimental condition. For example, we looked at Lin's concordance correlation coefficient (CCC) based on the average ratings of 20 behaviors between actor and observer perspectives. As a measure of agreement across experimental conditions, the CCC reflects how the observed data deviate from the line of

perfect concordance (Lin, 1989). The CCC values can range from -1 (perfect discordance) to +1 (perfect concordance) where a value of zero denotes the absence of any concordance/discordance. A CCC value greater than +0.95 is considered a substantial concordance (McBride, 2005).

We found substantial agreement between the actor and observer perspectives across both Study 1 (CCC = 0.99, 95% CI [0.96, 0.99]) and Study 2 (CCC = 0.99, 95% CI [0.99, 1.00]; see Figure 1). Similarly, we found substantial agreement between the definition conditions (Study 1: CCC = 0.99, 95% CI [0.96, 0.99]; Study 2: CCC = 0.99, 95% CI [0.96, 0.99]; see Figure 2).

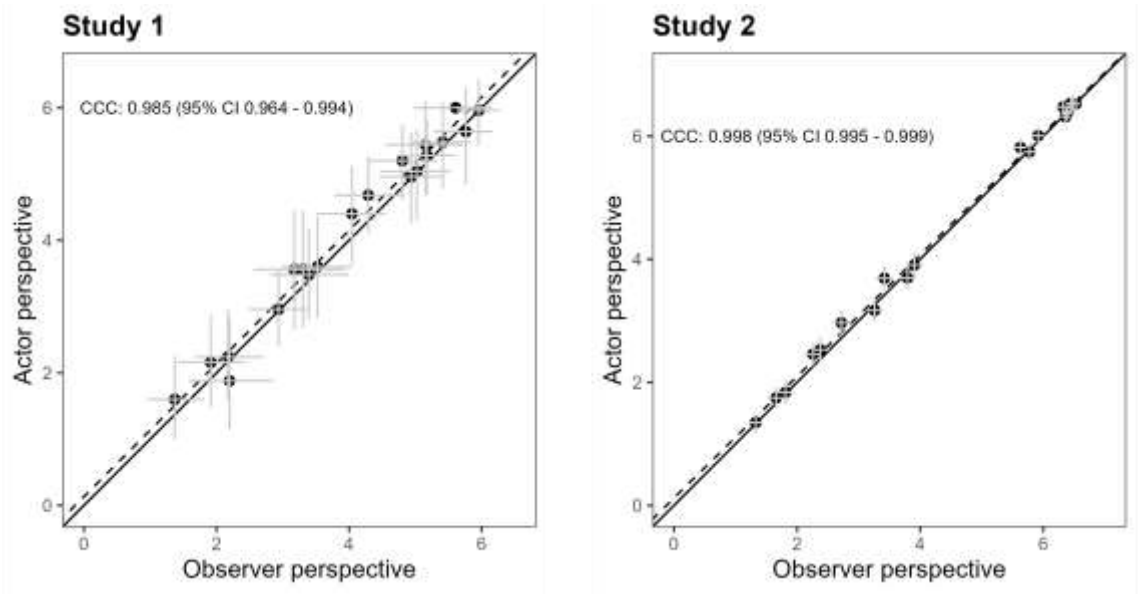


Figure 1. Lin's concordance correlation plot for average scores of intentionality across 20 items from actor and observer perspectives. The solid line represents the line of perfect concordance. The error bars on the data points indicate 95% bootstrapped confidence intervals.

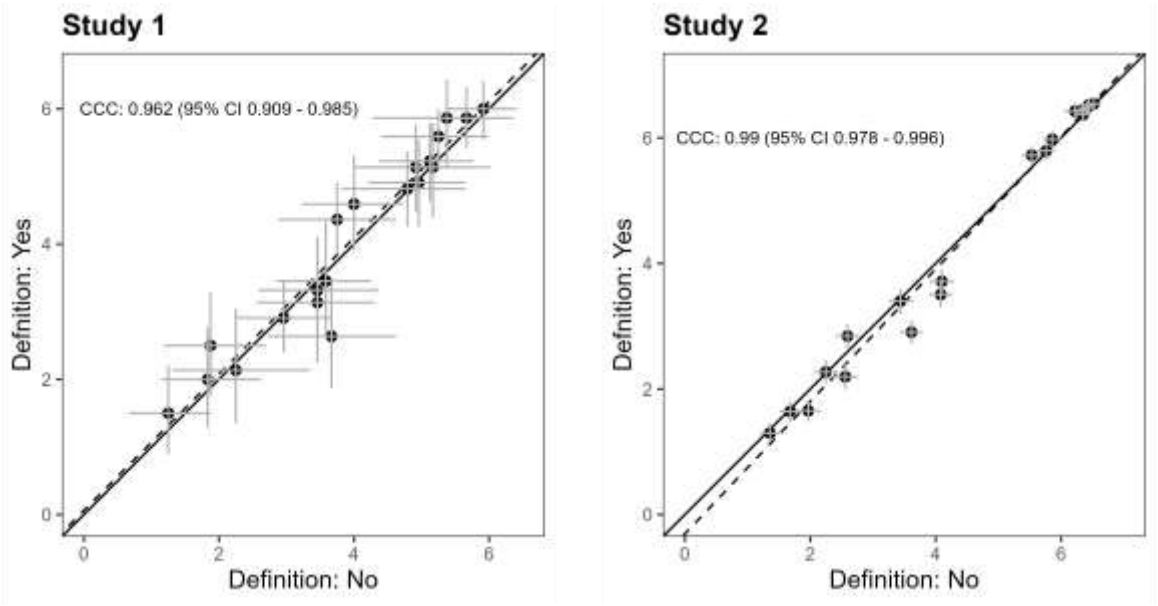


Figure 2. Lin's concordance correlation plot for average scores of intentionality with definition and without definition. The solid line represents the line of perfect concordance. The error bars on the data points indicate 95% bootstrapped confidence intervals.

Cross-sample comparison

We examined whether there were cross-cultural variations in intentionality ratings. The results indicate a relatively weak correlation between the samples of Study 1 and Study 2 (CCC = 0.24, 95% CI [-0.17, 0.59]). However, we found substantial agreement with a similar analysis between Study 2 and the original study responses (CCC = 0.97, 95% CI [0.93, 0.99]).

Actor-observer differences

We conducted two sets of 20 independent *t*-tests testing the differences in intentionality ratings across perspective and definition conditions. For detailed results, see Table S11–S14 in the supplemental material.

Mixed-model approach: An exploratory analysis

Although independent *t*-tests conducted in the previous section help test the differences in intentionality ratings across experimental conditions within each item, they are less helpful in summarizing the actor vs. observer or definition vs. without definition differences across 20 items. For example, a set of *t*-tests comparing actor-observer differences ignores the fact that the experiment also involved manipulation of definition.

Therefore, we conducted a linear mixed-effects analysis to draw summary conclusions based on within-subject responses to 20-items, between-subject experimental conditions, and data from the three studies (i.e., original study, Study 1, and Study 2). We pooled the responses across the three studies' samples into one to conduct a linear mixed-effects analysis—this practice is often termed integrative data analysis (IDA; Curran & Hussong, 2009). Linear mixed-effects models (LMEMs) have several advantages. LMEMs allows multiple data sets to be pooled into one while also taking into account variability within and across participants and items simultaneously.

We constructed the linear mixed model (using the *lme4* package in R; Bates et al., 2015) in R. The *p* values were calculated using the Satterthwaite approximation for the mixed-effects regressions (Kuznetsova et al., 2017). The mixed-effects regression analysis included perspective (observer vs. actor) and definition (with definition vs. without definition) conditions as fixed-effect predictors of intentionality ratings. Each participant rated intentionality across 20 described behaviors; therefore, the mixed-effect regression model carried three random

intercepts: participant's ID, item identifier, and sample identifier. We found no indication that a supplied definition was a fixed effect predictor of intentionality ratings. However, we found support for perspective as a fixed effect predictor (see Table 3). The results of linear mixed-effects models separately for each sample (original study, Study 1, and Study 2) are detailed in Table S17 (see Figure S2) of the supplementary material.

Table 3. The results of mixed-effects regression analysis with intentionality ratings as the dependent variable.

	Model 1				**Model 2**			
Predictors	**B**	**S.E.**	**CI**	***p***	**B**	**S.E.**	**CI**	***p***
(Intercept)	4.16	0.44	3.29 – 5.02	**<0.001**	3.50	0.96	1.62 – 5.38	**<0.001**
Perspective (observer vs. actor)	0.13	0.04	0.05 – 0.21	**0.001**	0.26	0.06	0.14 – 0.39	**<0.001**
Definition (with vs. without)	-0.06	0.04	-0.14 – 0.02	0.160	-0.06	0.04	-0.14 – 0.02	0.160
Valence (negative vs. non-negative)					0.82	1.06	-1.26 – 2.90	0.441
Perspective × valence					-0.16	0.06	-0.28 – -0.04	**0.007**
Random effects								
σ^2	2.71				2.71			
$\tau_{00\ \text{Participant ID}}$	0.26				0.26			
$\tau_{00\ \text{Item ID}}$	3.51				3.61			
$\tau_{00\ \text{Sample ID}}$	0.05				0.05			
ICC	0.58				0.59			
$N_{\text{Item ID}}$	20				20			
$N_{\text{Participant ID}}$	967				967			
N_{Sample}	3				3			
Observations	19338				19338			
Marginal R^2 / conditional R^2	0.001 / 0.585				0.014 / 0.596			

Note. Original study N = 104; Study 1 N = 46; Study 2 N= 817.

We further tested for the possibility that actor-observer differences were mainly or partly due to the responses to the negatively valenced behaviors. While answering the items with negative valence, study participants are likely motivated to deny they would do them intentionally. We identified four of the twenty items as clearly negative in valence: (i.e., "interrupted her mother," "ignored someone's arguments," "drove way above the speed limit," and "stole a pound of peaches") and the rest of the 16 items were classified as a non-negative valence category. We conducted a follow-up mixed-effect regression analysis with valence as a fixed effect predictor. We regressed intentionality ratings (DV) on perspective (observer vs. actor) and definition (with definition vs. without definition), valence (negative vs. non-negative), and the interaction term between valence and perspective as fixed-effect predictors. The participant's ID, item identifier, and sample identifier were included as random intercepts (see Model 2 of Table 3). We found some support that the actor-observer differences were moderated by the valence of the behaviors rated (see Figure 3). Specifically, actor-observer differences were larger with negatively valenced behaviors than with non-negative ones.

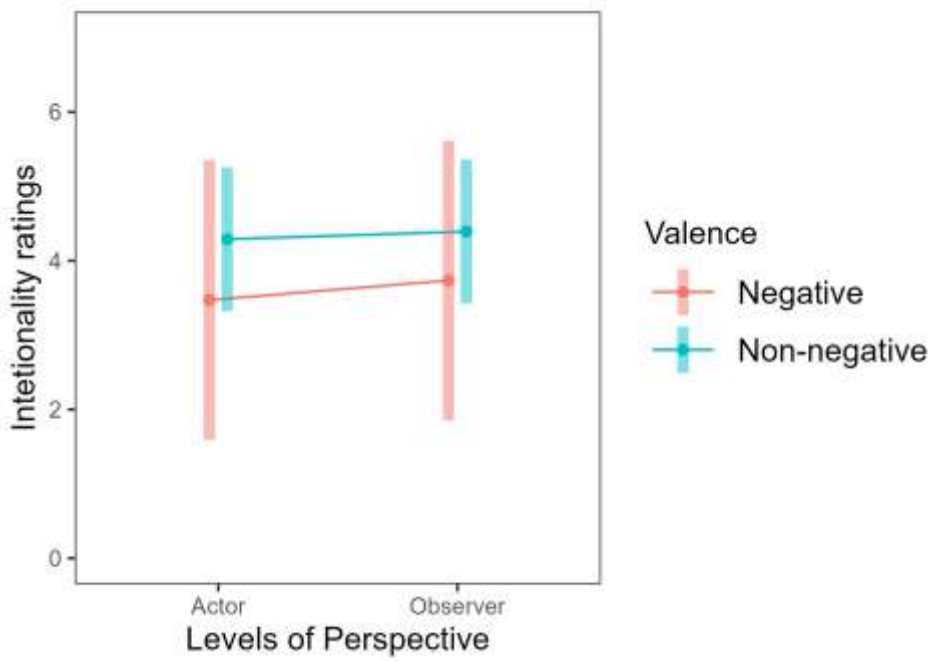


Figure 3. The interaction between perspective and valence on intentionality ratings.

Discussion

We aimed to conduct a close direct replication of Study 1 of Malle and Knobe (1997). Malle and Knobe (1997) argued for substantial agreement among intentional and unintentional behaviors as evidence for a shared concept of intentionality. Our replication results support the original study's central argument[4]. The presence or absence of the definition of intentionality had no bearing on the intentionality ratings.

We found support for the view that intentionality is a folk concept that people spontaneously and consistently attribute to behaviors. Across three different measures, we found a high level of agreement among the study participants across perspectives and definitions.

[4] Our main conclusions are based on the results of Study 2 because of the smaller sample size in Study 1.

We noted substantial agreement on intentionality ratings across actor-observer perspectives, with empirical support for actor-observer differences based on a mixed-effects regression analysis as part of an exploratory analysis. Specifically, we found that actors tended to attribute less intentionality than observers, with some evidence that valence partly explains these actor-observer differences.

Observations, limitations, and future directions

Our replications differed from the original studies in several ways. The stimuli used in the original article were tested with American undergraduates in the context of the 1990s. We used the same materials in an online survey with a more diverse population.

We noted some cross-sample variation based on the weak correlation between Study 1 and Study 2 samples' overall intentionality ratings. However, the small specialized sample we used in Study 1 does not allow for drawing any definitive conclusions.

Interestingly, we found support for actor-observer differences in intentionality attributions. We found that participants with the observer perspective attributed higher intentionality to a set of behaviors than participants with the actor perspective. However, the presence of actor-observer differences does not necessarily contradict the theoretical assertions of the folk-conceptual theory of behavior explanation (Malle & Knobe, 1997; Malle et al., 2007; Malle, 2004). The later theoretical and empirical works note possibilities of actor-observer asymmetries within the folk-conceptual theory (Malle et al., 2007). For example, observed actor-observer asymmetries in intentionality attributions might be due to differences in accessibility to

reasoning. For an actor, the reasons to act are readily accessible from memory, whereas the observer has no access to the decision-making processes of the actor.

Across four studies based on side-effects scenarios, Feltz and colleagues (2012) found support for actor-observer differences in intentionality attributions when study participants answered scenarios that described a harmful side-effect. In contrast, no such actor-observer differences emerged when scenarios described a helpful side-effect. The work by Feltz et al. (2012) highlights the role of self-serving bias on actor-observer differences noted in the literature. We tested whether self-serving bias contributed to the actor-observer differences found in our replication studies. The results suggest that self-serving bias may partly contribute to the actor-observer differences in intentionality attributions.

It is possible that we found actor-observer differences in the replication sample due to minor modifications to some of the behavior description statements—the target of actions in the original items (i.e., Ben, Mike, and Sue) was replaced with "someone." However, the results were the same when we conducted a robustness check based on replication Study 2. We excluded the responses to the modified items and tested for the actor-observer differences. The results support the actor-observer differences, and we reported the results of this sub-sample analysis in Table S18 of the supplementary material. The replication results are unlikely due to replacing the names of the targets with "someone."

We also note limitations that suggest promising directions for future research. Firstly, our Study 1 sample size was small, preventing us from drawing independent conclusions from the

findings. We directly address this shortcoming with a larger sample in Study 2, although participants were drawn from a White, Educated, Industrialized, Rich, and Democratic (WEIRD) society. Future work can further investigate and explore the folk concept of intentionality based on a larger study sample from some non-western societies with a possibility of testing for cross-cultural variations in intentionality attributions.

Secondly, we investigated the replicability based on the measures of agreement across participants. Although our replication results indicate strong support for the main predictions of the original study, we draw these conclusions in the absence of a formal framework that aids us in evaluating the replication based on agreement measures. For example, our conclusions do not compare the precision of the replication results relative to the original study. We currently lack an established framework to evaluate the replication results—both at individual-study and meta-analytic levels—based on statistical tests employed in the study. We recommend that the scholars focusing on psychology's meta-science develop a framework to evaluate replications based on agreement measures.

Open practices

All data, materials, analysis scripts, and preregistration documents are available at the Open Science Framework (https://osf.io/4q5ce/).

References

Ames, D. R., Knowles, E. D., Morris, M. W., Kalish, C. W., Rosati, A. D., & Gopnik, A. (2001). The social folk theorist: Insights from social and cultural psychology. In B. F., Malle, L. J., Moses, & D. A., Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 307-328). Cambridge, MA: MIT Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. DOI: https://doi.org/10.18637/jss.v067.i01

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: the simultaneous analysis of multiple data sets. *Psychological methods*, *14*(2), 81.

Dennison, S. M. (2007). Interpersonal relationships and stalking: Identifying when to intervene. *Law and Human Behavior, 31*, 353-367. Doi: 10.1007/s10979-006-9067-3

Feldman, G., Baumeister, R. F., & Wong, K. F. E. (2014). Free will is about choosing: The link between choice and the belief in free will. *Journal of Experimental Social Psychology*, *55*, 239-245.

Feltz, A., Harris, M., & Perez, A. (2012). Perspective in intentional action attribution. *Philosophical Psychology*, *25*(5), 673-687.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry*, *52*(5), 449-466.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley.

Jones, E. E., & Davis, K. E. (1965). From acts to dispositions: The attribution process in person perception. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 2, pp. 219–266). New York: Academic Press.

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology, 103*(1), 54–69. https://doi.org/10.1037/a0028347

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and brain sciences, 33*(4), 315-329.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389-402.

Lin, L. I. K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255-268.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433-442.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review*, *3*(1), 23-48.

Malle, B. F. (2004). *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press.

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology, 33,* 101-121. Doi: 10.1006/jesp.1996.1314

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2012). Moral, cognitive, and social: The nature of blame. *Social thinking and interpersonal behaviour*, 313-331.

Malle, B. F., Knobe, J. M., & Nelson, S. E. (2007). Actor-observer asymmetries in explanations of behavior: new answers to an old question. *Journal of personality and social psychology*, *93*(4), 491.

Malle, B. F., Knobe, J., O'Laughlin, M. J., Pearce, G. E., & Nelson, S. E. (2000). Conceptual structure and social functions of behavior explanations: Beyond person–situation attributions. *Journal of Personality and Social Psychology*, *79*(3), 309.

McBride, G. B. (2005). A proposal for strength-of-agreement criteria for Lin's concordance correlation coefficient. *NIWA client report: HAM2005-062*, *45*, 307-310.

Nosek, B. A., & Lakens, D. (2014). Registered reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, 45, 137-141. Doi: 10.1027/a000001

Ohtsubo, Y. (2007). Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect 1. *Japanese Psychological Research*, *49*, 100-110. Doi: 10.1111/j.1468-5884.2007.00337.x

Searle, J. R. (1995). The Construction of Social Reality. New York: Free Press.

Sloman, S. A., Fernbach, P. M., & Ewing, S. (2012). A causal model of intentionality judgment. *Mind & Language, 27*, 154-180. Doi: 10.1111/j.1468-0017.2012.01439.x

Swap, W. C. (1991). When prosocial behavior becomes altruistic: An attributional analysis. *Current Psychology: Research and Reviews*, 10, 49–64

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*(5), 675-691.

Vangelisti, A. L., & Young, S. L. (2000). When words hurt: The effects of perceived intentionality on interpersonal relationships. *Journal of Social and Personal Relationships, 17*, 393-424. Doi: 10.1177/0265407500173005

van't Veer, A.E.E, & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology, 67*, 2-12. DOI: 10.1016/j.jesp.2016.03.004

Zalla, T., & Leboyer, M. (2011). Judgment of intentionality and moral evaluation in individuals with high functioning autism. *Review of Philosophy and Psychology*, *2*(4), 681-698.

Revisiting the Folk Concept of Intentionality:

Replications of Malle and Knobe (1997)

Supplementary Materials

Disclosures

Procedure

The replication conducted with Hong Kong students was conducted as part of a large replication project, where students participated voluntarily in a survey where we attempted to replicate several findings from the judgment and decision-making literature.

Similarly, the replication conducted with Amazons Mechanical Turk (MTurk) workers was conducted as part of a large replication project, where MTurk workers were paid a nominal amount to participate in a survey where we attempted to replicate several findings from the judgment and decision-making literature.

Pre-registrations

Pre-registrations were conducted prior to data collection.

Data collection

Both data collections were combined with other independent replication attempts and displayed in a randomized order. Data collection was completed before conducting data analysis.

Conditions reporting

All collected conditions are reported.

Data exclusions

In the Hong Kong sample, the data collection was performed on student designing replications. We excluded participants who designed the study.

We note that the pre-registration plans included different references to possible exclusion criteria addressing seriousness, English proficiency, etc. We conducted our analyses both with and without exclusions, and found that exclusions had little effect on the results. The undergraduate samples and MTurk samples were proficient and serious.

Variable reporting

All variables collected for the replications are reported and included in the provided data.

Open science

Datasets, code, and supplementary materials were made available on the Open Science Framework at: https://osf.io/4q5ce/

Table S1. Deviations from pre-registration

Components of pre-registration	Were there deviations?	If yes describe the details of the deviation(s)	Rationale for deviation	Additional notes
Procedures	No	N/A	N/A	N/A
Power analysis	No	N/A	N/A	N/A
Exclusion rules	No	N/A	N/A	N/A
Evaluation criteria	Yes	During pre-registration we proposed to use the LeBel et al.'s (2018) for actor-observer differences, and definition (with vs. without) differences. LeBel et al's frame work is not helpful to evaluate agreement measures on intentionality ratings.	Main focus of the replication was the prediction: There will be a substantial agreement in judging behaviors for their intentionality.	We substantially focus on the agreement measure on intentionality ratings. The replication results are qualitatively evaluated.
Predictions	Yes	**Non-registered prediction:** Participants will show substantial agreement in judging behaviours for their intentionality.	The review process and feedback from the authors of original study we learnt that the prediction based on the inter-rater agreement on intentionality ratings was the main focus of the target study.	The evaluation of the replication (the success or failure to replicate) will be based on the results of the inter-rater agreement measures.

	Pre-registered predictions: *Actor perspective/ Observer perspective* Hypothesis 1: Results yield from the actor perspective will have a higher rating of intentionality of the behaviours than the observer perspective Hypothesis 2: Results yield from the observer perspective will have a lower rating of intentionality of the behaviours than the actor perspective *Definition provided/Definition not provided* Hypothesis 3: Results yield from the data with definition provided will have a higher rating of intentionality of the behaviours than the data without the definition provided Hypothesis 4: Results yield from the data with definition not provided will have a lower rating of intentionality of the behaviours than the data with definition provided	Evaluation of the pre-registered predictions will be based on mixed-effect regression analysis. The results based on T-tests will be reported in the Supplementary Materials document.	(please see 'Analysis' row-heading within this table for follow-up details)

Analysis	Yes	1) Pre-registration proposed t-test each of the 20 behavior items across perspective (actor vs. observer), and definition (with vs. without). Instead of two-sets of 20 t-tests, we conducted a mixed-effects regression (for each study, and integrative analysis combining responses from three samples: original study, study 1, and study 2).	mixed-effects regression is the most appropriate (study involved Perspective × Definition between-subjects manipulation) analysis given the data and allows for a single analysis to draw conclusions.	We draw conclusions based on the significance of the fixed effect predictors [perspective (actor vs. observer), and definition (with vs. without)].
		2) The original study reported two measures of inter-rater agreement: 1) Average inter-correlation; 2) Average rater whole correlation. Apart from these we also report two additional measures: 3) Intraclass Correlation Coefficient (ICC); 4) reliability coefficient (α measure).	We do not deviate from the original inter-rater agreement measures, but share results on additional measures.	N/A
Analysis: ANOVA	Yes	We decided to drop the pre-registered ANOVA analysis from main manuscript because we wanted to mainly focus on Inter-rater agreement on the intentionality judgments.	Our decision to drop the pre-registered ANOA analysis mainly stems from the valid critique that ANOVA analysis is not apt and does not serve the main	The pre-registered ANOVA (although flawed) are geared toward testing actor-observer asymmetry in intentionality judgment.

	Instead of the ANOVA analysis we report results based on an improved analytic strategy: linear mixed-effects models (LMEM). We report the results of LMEM as Exploratory data analysis.	aim of the original study. The aim of the study was to test the replicability of substantial Inter-rater agreement on the intentionality judgments	Please refer to the 'Pre-registration analysis plan' and 'Pre-registered analysis results' sections of the Supplementary material document to see the results of ANOVA analysis.

Full Materials

Instruction

Before the 20 behavioral statements were presented, participants in the actor perspective & with-definition condition read:

> "Please look at the 20 statements below. Each statement describes you doing something. Your task will be to rate whether you would do that *intentionally*. **What do we mean by intentional? This means that you had a reason to do what you did and that you chose to do so.**"

Participants in the actor perspective & without-definition condition read:

> "Please look at the 20 statements below. Each statement describes you doing something. Your task is to rate whether you would do that *intentionally*."

Participants in the observer perspective & with-definition condition read:

> "Please look at the 20 statements below. Each statement describes Anne doing something. Your task will be to rate whether Anne would do that *intentionally*. **What do we mean by intentional? This means that Anne had a reason to do what she did and that she chose to do so.**"

Participants in the observer perspective & without-definition condition read:

> "Please look at the 20 statements below. Each statement describes Anne doing something. Your task is to rate whether Anne would do that *intentionally*."

Comprehension question (only included in Replication Study 2)
To make sure you understood our definition, please answer the following comprehension question - What do we mean by intentional? This means that -

1. You had a reason to do what you did.
2. You chose to do what you did.
3. You had a reason to do what you did and that you chose to do so.

Table S2.

Twenty Behavioral statements. Each statement is rated on a 8-point Likert scale from "Not at all intentional" (0) to "Completely intentional" (7).

	Actor Perspective	Observer Perspective
Item1	I am sweating.	Anne is sweating.
Item2	I was yawning during the lecture.	Anne was yawning during the lecture.
Item3	I was grinding my teeth during the test.	Anne was grinding her teeth during the test.
Item4	I had a craving for cherries after dinner.	Anne had a craving for cherries after dinner.
Item5	I believed that I had the flu.	Anne believed that she had the flu.
Item6	I won a prize in the lottery.	Anne won a prize in the lottery.
Item7	I am in a great mood today.	Anne is in a great mood today.
Item8	I am infatuated with someone.	Anne is infatuated with someone.
Item9	I was worrying about the test results.	Anne was worrying about the test results.
Item10	I got admitted to a university.	Anne got admitted to a university.
Item11	I interrupted my mother.	Anne interrupted her mother.
Item12	I ignored someone's arguments.	Anne ignored someone's arguments.
Item13	I drove way above the speed limit.	Anne drove way above the speed limit.
Item14	I applauded the musicians.	Anne applauded the musicians.
Item15	I greeted my uncle politely.	Anne greeted her uncle politely.
Item16	I refused the salesman's offer.	Anne refused the salesman's offer.
Item17	I stole a pound of peaches.	Anne stole a pound of peaches.
Item18	I asked someone out for dinner.	Anne asked someone out for dinner.
Item19	I invited someone to have lunch with me.	Anne invited someone to have lunch with her.
Item20	I watered my new plants.	Anne watered her new plants.

Exclusion questions (only included in Replication Study 2)

- How is your English proficiency?
  Answered on a 7-point Likert scale from "Very Bad" = (1) to "Very Good" = (7).
- How serious were you in taking this questionnaire?
  Answered on a 5-point Likert scale from "Not at all" = (1) to "Very much" = (5).

Pre-registration analysis plan

Based on the results of Malle and Knobe (1997), we did not expect to find support for any differences between perspective conditions (actor vs. observer) or definition conditions (with vs. without) on intentionality ratings. Framed in terms of expected effects, we expected to observe effect sizes ranging $-0.2 < d < 0.2$ for both perspective and definition on intentionality (Cohen, 1988). We summarized the hypotheses in the present study in Table S3.

Table S3. Summary of pre-registered predictions.

Perspective (actor vs. observer)
Hypothesis 1a: The actor perspective leads to a higher average rating of intentionality than the observer perspective.
Hypothesis 1b: The actor perspective leads to a lower average rating of intentionality than the observer perspective.
Definition (with vs. without)
Hypothesis 2a: Presentation of a definition of intentionality leads to a higher average rating of intentionality relative to the condition in which no definition is given.
Hypothesis 2b: Presentation of a definition of intentionality leads to a lower average rating of intentionality relative to the condition in which no definition is given.

Power analysis

We used G*Power (Faul et al., 2007) to estimate the effect sizes of the perspective and definition conditions on intentionality rating in the original study. As Cohen's *d* is below .20 for small effects (Cohen, 1988), to obtain power = .80 and $\alpha = .05$, a sample size of 788 was required. We conducted Study 2 with a well-powered sample size. For Study 1, however, the sample size was determined by the availability of sample (see the details below), hence we considered Study 1 as a pre-test.

Analysis strategy

We report both methods resembling the original (i.e., repeating the same analytic strategy as in Malle & Knobe, 1997), as well as an improved analytic approach. Our pre-registered analytic approach included two sets of 20 pairwise independent t-tests, testing the differences in intentionality ratings across different perspective and definition conditions. Furthermore, as part of the pre-registration we proposed to conduct a one-way analysis of variance (ANOVA) to examine the effects of perspective and definition, respectively, on intentionality rating[1]. We calculated the average internationality score for each participant by averaging their responses across the twenty descriptions of the behavior. Based on these average scores per participant, we performed the one-way ANOVA to examine the effects of the perspective and definition conditions on intentionality rating.

Pre-registered analysis results

Descriptive statistics of the measures across both of Study 1 and 2 are noted in Table S5–S8 of the supplementary material. Two sets of 20 independent t-tests, testing the differences in intentionality ratings across perspective and definition conditions, revealed not support for differences in ratings between perspectives or the presence (vs. absence) of intentionality definition. Malle and Knobe (1997) thus concluded that the findings provided support to a shared folk concept of intentionality. For details results see Table S11–S14 in the supplement.

We summarized the results of the ANOVA in Table S4. The results testing the differences in perspective (actor vs. observer) found no support in Study 1 (Study 1: $d = -0.60$, 95% *CI* = [-1.21, 0.01], $BF_{10} = 1.49$), likely due to the small sample. However, in Study 2, consistent with Hypothesis 1b we found intentionality ratings were lower in the actor perspective than the observer perspective (Study 2: $d = -0.21$, 95% *CI* = [-0.34, - 0.07], $BF_{10} = 9.66$). For the provision of definition, we found no support for differences in intentionality ratings between with-definition and without-definition conditions (Study 1: $d = -0.17$, 95% *CI* = [-0.77, 0.43], $BF_{10} = 0.75$; Study 2: $d = 0.10$, 95% *CI* = [-0.04, 0.24], $BF_{10} = 0.89$).

[1] Study 1 of the original paper did not report ANOVA analysis. Authors mainly presented reliability/agreement analysis.

Table S4.

One-way ANOVA of perspective and definition on intentionality rating.

	Comparisons	*df*1	*df*2	*F*	*p*	Cohen's *d*	Bayes Factor
Original study	Perspective (actor vs. observer)	1	102	0.24	0.628	-0.10 [-0.52, 0.31]	BF_{10} = 0.66; BF_{01}=1.51
	Definition (with vs. without)	1	102	0.11	0.739	0.07 [-0.32, 0.45]	BF_{10} = 0.63; BF_{01}=1.59
Study 1	Perspective (actor vs. observer)	1	44	4.07	0.049	-0.60 [-1.21, 0.01]	BF_{10} = 1.47; BF_{01}=0.68
	Definition (with vs. without)	1	44	0.34	0.565	-0.17 [-0.77, 0.43]	BF_{10} = 0.34; BF_{01}=2.98
Study 2	Perspective (actor vs. observer)	1	815	8.69	0.003	-0.21 [-0.34, -0.07]	BF_{10} = 9.66; BF_{01}=0.10
	Definition (with vs. without)	1	815	2.14	0.144	0.10 [-0.04, 0.24]	BF_{10} = 0.89; BF_{01}=1.12

Note. For Bayesian analysis, the prior (width parameter) was set at 0.1.

Supplementary Analyses

(not reported in the main manuscript)

Descriptive statistics

Table S5.

Intentionality ratings.

Item No.	Original Study M (S.D) N= 104	Study 1 M (S.D) N= 46	Study 2 M (S.D) N= 817
Item 1	6.39 (1.07)	1.37 (1.50)	6.45 (1.18)
Item 2	2.70 (1.98)	1.91 (1.74)	3.26 (1.95)
Item 3	1.41 (1.60)	2.17 (1.80)	1.66 (1.74)
Item 4	3.20 (1.95)	3.30 (2.13)	3.42 (2.06)
Item 5	6.40(1.08)	3.17 (2.13)	6.48 (1.19)
Item 6	0.94 (1.52)	2.20 (2.22)	1.81 (1.96)
Item 7	2.69 (1.99)	2.93 (1.51)	2.38 (2.09)
Item 8	6.53 (0.84)	3.52 (1.83)	6.53 (1.11)
Item 9	3.69 (1.91)	3.39 (2.02)	3.91 (2.05)
Item 10	5.37 (1.67)	5.02 (1.82)	5.92 (1.47)
Item 11	2.00 (1.63)	4.04 (1.79)	2.72 (2.03)
Item 12	6.36 (1.19)	4.29 (1.73)	6.32 (1.59)
Item 13	3.78 (2.19)	4.93 (1.64)	3.79 (2.01)
Item 14	1.38 (1.61)	5.17 (1.52)	1.33 (1.67)
Item 15	5.94 (1.31)	4.80 (1.81)	6.36 (1.18)
Item 16	4.58 (1.94)	5.41 (1.53)	5.63 (1.58)
Item 17	2.23 (1.72)	5.61 (1.95)	2.26 (1.94)
Item 18	6.22 (1.28)	5.96 (1.13)	6.40 (1.14)
Item 19	5.77 (1.49)	5.76 (1.51)	6.38 (1.17)
Item 20	5.22 (1.73)	5.15 (1.95)	5.78 (1.45)

Note. Mean and S.D. based on responses across conditions.

Table S6.

Original study: Intentionality ratings by perspective and definition.

	Original Study							
	Perspective				Definition			
Item list	Actor (n = 32)		Observer (n = 72)		With (n = 44)		Without (n = 60)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Item1	6.31	1.15	6.43	1.05	6.23	1.29	6.52	0.87
Item2	3.06	2.08	2.54	1.93	2.68	1.89	2.72	2.06
Item3	1.39	1.71	1.42	1.56	1.43	1.80	1.39	1.45
Item4	2.72	1.92	3.42	1.94	3.00	1.90	3.35	1.99
Item5	6.25	1.39	6.47	0.92	6.32	1.16	6.47	1.03
Item6	0.84	1.63	0.99	1.48	0.91	1.49	0.97	1.55
Item7	3.00	2.24	2.56	1.87	2.70	1.98	2.68	2.01
Item8	6.44	1.13	6.57	0.67	6.57	0.70	6.50	0.93
Item9	3.56	1.93	3.75	1.91	3.77	1.92	3.63	1.92
Item10	5.38	1.64	5.36	1.69	5.39	1.45	5.35	1.82
Item11	2.13	2.04	1.94	1.41	1.93	1.42	2.05	1.77
Item12	6.31	1.42	6.38	1.08	6.48	1.11	6.27	1.25
Item13	3.69	2.15	3.82	2.23	3.86	2.28	3.72	2.15
Item14	1.06	1.27	1.51	1.74	1.66	1.87	1.17	1.38
Item15	6.19	1.20	5.83	1.35	5.91	1.29	5.97	1.34
Item16	4.38	1.93	4.67	1.95	4.39	2.00	4.72	1.90
Item17	2.13	1.84	2.28	1.67	2.00	1.57	2.40	1.82
Item18	6.53	0.67	6.08	1.46	6.27	1.06	6.18	1.43
Item19	5.56	1.68	5.86	1.40	5.75	1.46	5.78	1.52
Item20	4.91	1.91	5.36	1.64	5.09	1.76	5.32	1.71

Table S7.

Study 1: *Intentionality ratings by perspective and definition.*

	Perspective				**Definition**			
Item list	**Actor (n = 21)**		**Observer (n = 25)**		**With (n = 22)**		**Without (n = 24)**	
	Mean	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**	**Mean**	**S.D.**
Item1	1.10	1.34	1.60	1.61	**1.5**	**1.54**	**1.25**	**1.48**
Item2	1.62	1.72	2.16	1.75	2.00	1.72	1.83	1.79
Item3	2.10	1.95	2.24	1.71	2.50	1.77	1.88	1.83
Item4	3.00	2.00	3.56	2.24	3.14	2.14	3.46	2.15
Item5	2.71	1.93	3.56	2.26	2.64	1.81	3.67	2.32
Item6	2.57	2.46	1.88	1.99	2.14	1.93	2.25	2.49
Item7	2.90	1.61	2.96	1.46	2.91	1.23	2.96	1.76
Item8	3.43	1.75	3.60	1.94	3.45	1.99	3.58	1.72
Item9	3.29	2.35	3.48	1.73	3.32	1.86	3.46	2.19
Item10	5.00	2.02	5.04	1.67	5.14	1.52	4.92	2.08
Item11	3.62	1.66	4.40	1.85	4.36	1.26	3.75	2.15
Item12	3.80	1.91	4.68	1.49	4.59	1.65	4.00	1.78
Item13	4.90	1.61	4.96	1.70	4.91	1.54	4.96	1.76
Item14	5.05	1.63	5.28	1.46	5.23	1.38	5.13	1.68
Item15	4.33	2.15	5.20	1.38	4.82	1.30	4.79	2.21
Item16	5.33	1.46	5.48	1.61	5.59	1.01	5.25	1.89
Item17	5.14	2.31	6.00	1.53	5.86	1.42	5.38	2.34
Item18	5.95	1.02	5.96	1.24	6.00	1.02	5.92	1.25
Item19	5.90	1.18	5.64	1.75	5.86	1.08	5.67	1.83
Item20	4.81	2.14	5.44	1.78	5.14	1.67	5.17	2.22

Table S8.

Study 2: *Intentionality ratings by perspective and definition.*

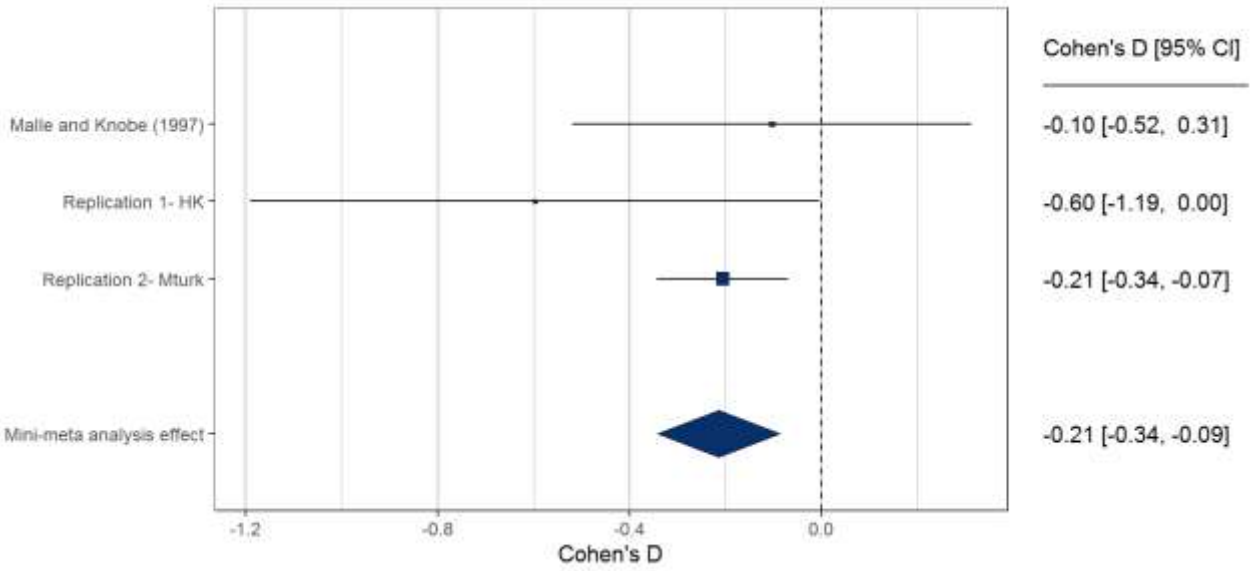
	Replication Study 2							
	Perspective				Definition			
Item list	Actor (n = 405)		Observer (n = 412)		With (n = 408)		Without (n = 409)	
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
Item1	**6.36**	**1.25**	**6.53**	**1.1**	**6.48**	**1.08**	**6.42**	**1.27**
Item2	3.35	1.99	3.17	1.90	2.90	1.95	3.61	1.87
Item3	1.57	1.75	1.75	1.72	1.64	1.67	1.68	1.80
Item4	3.14	2.10	3.69	1.98	3.40	2.05	3.44	2.06
Item5	6.44	1.19	6.52	1.19	6.52	1.13	6.44	1.25
Item6	1.78	1.92	1.84	2.00	1.65	1.85	1.97	2.05
Item7	2.22	2.07	2.53	2.10	2.20	2.09	2.56	2.08
Item8	6.54	1.10	6.53	1.13	6.54	1.03	6.52	1.19
Item9	3.90	2.05	3.91	2.05	3.71	2.00	4.10	2.07
Item10	5.82	1.55	6.01	1.39	5.98	1.46	5.86	1.48
Item11	2.46	1.99	2.97	2.05	2.84	1.99	2.59	2.08
Item12	6.17	1.84	6.48	1.27	6.42	1.37	6.22	1.77
Item13	3.89	2.03	3.70	1.98	3.50	2.04	4.08	1.93
Item14	1.31	1.66	1.35	1.69	1.30	1.57	1.35	1.76
Item15	6.40	1.12	6.32	1.23	6.37	1.12	6.35	1.24
Item16	5.43	1.70	5.82	1.44	5.72	1.50	5.53	1.66
Item17	2.06	1.87	2.46	1.99	2.27	1.92	2.25	1.96
Item18	6.39	1.16	6.40	1.13	6.45	1.09	6.35	1.20
Item19	6.40	1.15	6.37	1.20	6.42	1.10	6.35	1.24
Item20	5.81	1.36	5.74	1.54	5.79	1.39	5.76	1.51

Mini meta-analysis[2]

We summarized the three the samples (original study, Study 1, and Study 2) using a mini meta-analysis (see Figure S1) (Goh, Hall, & Rosenthal, 2016; Lakens & Etz, 2017). We ran the mini-meta with random effects model using the DerSimonian-Laird method using the *metafor* package in R (Viechtbauer, 2010). We did not find support the predictions based on provision of the definition (Participant level analysis: d =0.09 [-0.04, 0.21]). In testing the predictions based on perspectives, the effect size estimation based on participant level analysis indicated support for Hypothesis 1b (d = -0.21 [-0.34, -0.09]).

Figure S1.

Forest plot of the mini meta-analyses. Participant level analysis of self-other asymmetry on intentionality attribution.



[2] Please refer to minimeta.Rmd file available at https://osf.io/xd5ak/?view_only=ee154c8295bd4415a6a20084829a06a7 to reproduce the results of the mini-meta analysis.

Correlations between the Original Study and the Replication Studies

Actor Perspective:
Malle & Knobe (1997) & Study 1: $r(19) = 0.18$, $p = .441$.
Malle & Knobe (1997) & Study 2: $r(19) = 0.98$, $p < .001$.
Studies 1 & 2: $r(19) = 0.23$, $p = .328$.

Observer Perspective:
Malle & Knobe (1997) & Study 1: $r(19) = 0.28$, $p = .236$.
Malle & Knobe (1997) & Study 2: $r(19) = 0.98$, $p < .001$.
Studies 1 & 2: $r(19) = 0.29$, $p = .213$.

With Definition:
Malle & Knobe (1997) & Study 1: $r(19) = 0.21$, $p = .378$.
Malle & Knobe (1997) & Study 2: $r(19) = 0.97$, $p < .001$.
Studies 1 & 2: $r(19) = 0.247$, $p = .293$.

Without Definition:
Malle & Knobe (1997) & Study 1: $r(19) = 0.27$, $p = .249$.
Malle & Knobe (1997) & Study 2: $r(19) = 0.99$, $p < .001$.
Studies 1 & 2: $r(19) = 0.28$, $p = .237$.

As above shown, intentionality ratings in the same condition demonstrated strong correlations between the original study and Study 2.

Independent t-tests

Table S9.

Original study: Item by item analysis of Actor-Observer asymmetry on intentionality ratings (N = 104).

Item code	Mean (actor)	Mean (observer)	Independent *t*-test	Cohen d with 95% CI
ACT1	6.31	6.43	t(54.82) = -0.5; p = 0.621	-0.11 [-0.52, 0.31]
ACT2	3.06	2.54	t(55.69) = 1.21; p = 0.233	0.26 [-0.16, 0.67]
ACT3	1.39	1.42	t(52.69) = -0.08; p = 0.934	-0.02 [-0.43, 0.40]
ACT4	2.72	3.42	t(60.09) = -1.7; p = 0.094	-0.36 [-0.78, 0.06]
ACT5	6.25	6.47	t(43.47) = -0.83; p = 0.413	-0.19 [-0.59, 0.24]
ACT6	0.84	0.99	t(54.64) = -0.42; p = 0.674	-0.09 [-0.51, 0.33]
ACT7	3.00	2.56	t(50.94) = 0.98; p = 0.332	0.22 [-0.21, 0.63]
ACT8	6.44	6.57	t(40.86) = -0.61; p = 0.544	-0.14 [-0.55, 0.29]
ACT9	3.56	3.75	t(58.94) = -0.46; p = 0.649	-0.10 [-0.51, 0.32]
ACT10	5.38	5.36	t(61.15) = 0.04; p = 0.969	0.01 [-0.41, 0.42]
ACT11	2.13	1.94	t(44.69) = 0.45; p = 0.652	0.10 [-0.32, 0.51]
ACT12	6.31	6.38	t(47.53) = -0.22; p = 0.826	-0.05 [-0.46, 0.37]
ACT13	3.69	3.82	t(61.63) = -0.29; p = 0.776	-0.06 [-0.48, 0.36]
ACT14	1.06	1.51	t(79.92) = -1.49; p = 0.141	-0.30 [-0.73, 0.10]
ACT15	6.19	5.83	t(66.48) = 1.33; p = 0.187	0.28 [-0.14, 0.70]
ACT16	4.38	4.67	t(60.1) = -0.71; p = 0.481	-0.15 [-0.57, 0.27]
ACT17	2.13	2.28	t(54.59) = -0.4; p = 0.69	-0.09 [-0.5, 0.33]
ACT18	6.53	6.08	t(101.84) = 2.14; p = 0.035	0.39 [0.03, 0.88]
ACT19	5.56	5.86	t(50.8) = -0.88; p = 0.384	-0.19 [-0.60, 0.23]
ACT20	4.91	5.36	t(52.23) = -1.17; p = 0.247	-0.26 [-0.67, 0.17]

Note: Independent t-test results

Table S10.

Original study: Item by item analysis of without vs. with definition asymmetry on intentionality ratings (N = 104).

Item code	Mean (without definition)	Mean (with definition)	Independent *t*-test	Cohen's d with 95% CI
ACT1	6.52	6.23	t(70.86) = 1.29; p =0.203	0.26 [-0.14, 0.65]
ACT2	2.72	2.68	t(96.93) = 0.09; p =0.929	0.02 [-0.37, 0.41]
ACT3	1.39	1.43	t(80.81) = -0.13; p =0.899	-0.03 [-0.42, 0.37]
ACT4	3.35	3.00	t(95.02) = 0.91; p =0.366	0.18 [-0.21, 0.57]
ACT5	6.47	6.32	t(86.44) = 0.68; p =0.501	0.14 [-0.26, 0.52]
ACT6	0.97	0.91	t(94.8) = 0.19; p =0.849	0.04 [-0.35, 0.43]
ACT7	2.68	2.70	t(93.57) = -0.05; p =0.957	-0.01 [-0.4, 0.38]
ACT8	6.50	6.57	t(101.95) = -0.43; p =0.67	-0.08 [-0.47, 0.30]
ACT9	3.63	3.77	t(93) = -0.37; p =0.715	-0.07 [-0.46, 0.32]
ACT10	5.35	5.39	t(101.26) = -0.11; p =0.91	-0.02 [-0.41, 0.37]
ACT11	2.05	1.93	t(101.12) = 0.38; p =0.707	0.07 [-0.31, 0.46]
ACT12	6.27	6.48	t(98.15) = -0.91; p =0.367	-0.18 [-0.57, 0.21]
ACT13	3.72	3.86	t(89.59) = -0.33; p =0.74	-0.07 [-0.46, 0.32]
ACT14	1.17	1.66	t(75.52) = -1.48; p =0.143	-0.30 [-0.68, 0.10]
ACT15	5.97	5.91	t(94.71) = 0.22; p =0.825	0.04 [-0.35, 0.43]
ACT16	4.72	4.39	t(89.84) = 0.85; p =0.398	0.17 [-0.22, 0.56]
ACT17	2.40	2.00	t(99.17) = 1.2; p =0.233	0.24 [-0.15, 0.63]
ACT18	6.18	6.27	t(101.97) = -0.37; p =0.716	-0.07 [-0.46, 0.32]
ACT19	5.78	5.75	t(94.67) = 0.11; p =0.91	0.02 [-0.37, 0.41]
ACT20	5.32	5.09	t(91.25) = 0.65; p =0.515	0.13 [-0.26, 0.52]

Note: Independent t-test results

Table S11.

Study 1: Item by item analysis of actor-observer asymmetry on intentionality ratings (N = 46).

Item code	Mean (actor)	Mean (observer)	t-test	Cohen's d with 95% CI
Item1	1.10	1.60	t(44)=-1.16; p =0.251	-0.34 [-0.93, 0.24]
Item2	1.62	2.16	t(42.90)=-1.06; p =0.297	-0.31 [-0.89, 0.27]
Item3	2.10	2.24	t(40.30)=-0.27; p =0.792	-0.08 [-0.66, 0.50]
Item4	3.00	3.56	t(43.81)=-0.90; p =0.375	-0.26 [-0.85, 0.32]
Item5	2.71	3.56	t(43.98)=-1.37; p =0.177	-0.40 [-0.99, 0.18]
Item6	2.57	1.88	t(38.31)=1.04; p =0.307	0.31 [-0.28, 0.89]
Item7	2.90	2.96	t(40.87)=-0.12; p =0.904	-0.04 [-0.62, 0.54]
Item8	3.43	3.60	t(43.75)=-0.32; p =0.754	-0.09 [-0.67, 0.49]
Item9	3.29	3.48	t(36.20)=-0.31; p =0.755	-0.09 [-0.67, 0.49]
Item10	5.00	5.04	t(38.83)=-0.07; p =0.943	-0.02 [-0.60, 0.56]
Item11	3.62	4.40	t(43.79)=-1.51; p =0.138	-0.44 [-1.03, 0.14]
Item12	3.80	4.68	t(35.42)=-1.69; p =0.100	-0.51 [-1.10, 0.08]
Item13	4.90	4.96	t(43.30)=-0.11; p =0.910	-0.03 [-0.61, 0.55]
Item14	5.05	5.28	t(40.67)=-0.51; p =0.616	-0.15 [-0.73, 0.43]
Item15	4.33	5.20	t(32.99)=-1.59; p =0.121	-0.48 [-1.06, 0.12]
Item16	5.33	5.48	t(43.71)=-0.32; p =0.748	-0.10 [-0.68, 0.49]
Item17	5.14	6.00	t(33.63)=-1.45; p =0.155	-0.44 [-1.02, 0.16]
Item18	5.95	5.96	t(43.99)=-0.02; p =0.982	-0.01 [-0.58, 0.57]
Item19	5.90	5.64	t(42.14)=0.61; p =0.546	0.18 [-0.40, 0.76]
Item20	4.81	5.44	t(39.08)=-1.07; p =0.289	-0.32 [-0.90, 0.27]

Note: Independent t-test results

Table S12.

Study 1: Item by item analysis of without vs. with definition asymmetry on intentionality ratings (N = 46).

Item code	Mean (without definition)	Mean (with definition)	Independent *t*-test	Cohen's d with 95% CI
Item1	1.25	1.50	t(43.33) = -0.56; p =0.578	-0.17 [-0.74, 0.41]
Item2	1.83	2.00	t(43.89) = -0.32; p =0.749	-0.10 [-0.67, 0.48]
Item3	1.88	2.50	t(43.86) = -1.18; p =0.244	-0.35 [-0.93, 0.24]
Item4	3.46	3.14	t(43.66) = 0.51; p =0.614	0.15 [-0.43, 0.73]
Item5	3.67	2.64	t(43) = 1.69; p =0.099	0.5 [-0.09, 1.08]
Item6	2.25	2.14	t(42.9) = 0.17; p =0.863	0.05 [-0.53, 0.63]
Item7	2.96	2.91	t(41.27) = 0.11; p =0.912	0.03 [-0.55, 0.61]
Item8	3.58	3.45	t(41.68) = 0.23; p =0.816	0.07 [-0.51, 0.65]
Item9	3.46	3.32	t(43.78) = 0.23; p =0.816	0.07 [-0.51, 0.65]
Item10	4.92	5.14	t(41.98) = -0.41; p =0.683	-0.12 [-0.7, 0.46]
Item11	3.75	4.36	t(37.58) = -1.19; p =0.24	-0.35 [-0.93, 0.23]
Item12	4.00	4.59	t(42.96) = -1.15; p =0.255	-0.34 [-0.93, 0.25]
Item13	4.96	4.91	t(43.92) = 0.1; p =0.92	0.03 [-0.55, 0.61]
Item14	5.13	5.23	t(43.51) = -0.23; p =0.822	-0.07 [-0.65, 0.51]
Item15	4.79	4.82	t(37.72) = -0.05; p =0.96	-0.01 [-0.59, 0.56]
Item16	5.25	5.59	t(35.67) = -0.77; p =0.446	-0.22 [-0.81, 0.35]
Item17	5.38	5.86	t(38.49) = -0.86; p =0.393	-0.25 [-0.83, 0.33]
Item18	5.92	6.00	t(43.49) = -0.25; p =0.805	-0.07 [-0.65, 0.51]
Item19	5.67	5.86	t(37.82) = -0.45; p =0.657	-0.13 [-0.71, 0.45]
Item20	5.17	5.14	t(42.45) = 0.05; p =0.958	0.02 [-0.56, 0.59]

Table S13.

Study 2: Item by item analysis of actor-observer asymmetry on intentionality ratings (N = 817).

Item code	Mean (actor)	Mean (observer)	Independent t-test	Cohen's d with 95% CI
Item1	6.36	6.53	t(797.45)=-2.07; p =0.039	-0.15 [-0.28, -0.01]
Item2	3.35	3.17	t(812.06)=1.33; p =0.185	0.09 [-0.04, 0.23]
Item3	1.57	1.75	t(814.03)=-1.46; p =0.145	-0.10 [-0.24, 0.04]
Item4	3.14	3.69	t(809.93)=-3.86; p < .001	-0.27 [-0.41, -0.13]
Item5	6.44	6.52	t(814.74)=-1.02; p =0.31	-0.07 [-0.21, 0.07]
Item6	1.78	1.84	t(814.49)=-0.4; p =0.690	-0.03 [-0.17, 0.11]
Item7	2.22	2.53	t(815)=-2.10; p =0.036	-0.15 [-0.28, -0.01]
Item8	6.54	6.53	t(815)=0.12; p =0.907	0.01 [-0.13, 0.15]
Item9	3.90	3.91	t(814.81)=-0.03; p =0.977	0.00 [-0.14, 0.14]
Item10	5.82	6.01	t(803.1)=-1.77; p =0.077	-0.12 [-0.26, 0.01]
Item11	2.46	2.97	t(814.81)=-3.57; p < .001	-0.25 [-0.39, -0.11]
Item12	6.17	6.48	t(716.68)=-2.78; p =0.006	-0.19 [-0.33, -0.06]
Item13	3.89	3.70	t(813.79)=1.39; p =0.166	0.10 [-0.04, 0.23]
Item14	1.31	1.35	t(815)=-0.35; p =0.727	-0.02 [-0.16, 0.11]
Item15	6.40	6.32	t(810.31)=1.08; p =0.279	0.08 [-0.06, 0.21]
Item16	5.43	5.82	t(789.14)=-3.46; p =0.001	-0.24 [-0.38, -0.10]
Item17	2.06	2.46	t(813.52)=-2.94; p =0.003	-0.21 [-0.34, -0.07]
Item18	6.39	6.40	t(812.99)=-0.10; p =0.922	-0.01 [-0.14, 0.13]
Item19	6.40	6.37	t(814.33)=0.32; p =0.750	0.02 [-0.11, 0.16]
Item20	5.81	5.74	t(806.73)=0.66; p =0.509	0.05 [-0.09, 0.18]

Note: Independent t-test results

Table S14.

Study 2: Item by item analysis of without vs. with definition asymmetry on intentionality ratings (N = 817).

Item code	Mean (without definition)	Mean (with definition)	Independent *t*-test	Cohen's d with 95% CI
Item1	6.42	6.48	t(795.65) = -0.81; p =0.416	-0.06 [-0.19, 0.08]
Item2	3.61	2.90	t(813.43) = 5.3; p < .001	0.37 [0.23, 0.51]
Item3	1.68	1.64	t(810.75) = 0.29; p =0.773	0.02 [-0.12, 0.16]
Item4	3.44	3.40	t(814.99) = 0.26; p =0.791	0.02 [-0.12, 0.16]
Item5	6.44	6.52	t(807.89) = -0.95; p =0.341	-0.07 [-0.2, 0.07]
Item6	1.97	1.65	t(806.82) = 2.28; p =0.023	0.16 [0.02, 0.3]
Item7	2.56	2.20	t(814.89) = 2.48; p =0.013	0.17 [0.04, 0.31]
Item8	6.52	6.54	t(799.93) = -0.27; p =0.789	-0.02 [-0.16, 0.12]
Item9	4.10	3.71	t(814.11) = 2.7; p =0.007	0.19 [0.05, 0.33]
Item10	5.86	5.98	t(814.92) = -1.14; p =0.255	-0.08 [-0.22, 0.06]
Item11	2.59	2.84	t(813.5) = -1.75; p =0.080	-0.12 [-0.26, 0.01]
Item12	6.22	6.42	t(767.01) = -1.82; p =0.069	-0.13 [-0.26, 0.01]
Item13	4.08	3.50	t(812.24) = 4.14; p < .001	0.29 [0.15, 0.43]
Item14	1.35	1.30	t(805.17) = 0.47; p =0.635	0.03 [-0.1, 0.17]
Item15	6.35	6.37	t(807.47) = -0.25; p =0.804	-0.02 [-0.15, 0.12]
Item16	5.53	5.72	t(807.46) = -1.74; p =0.082	-0.12 [-0.26, 0.02]
Item17	2.25	2.27	t(814.83) = -0.18; p =0.853	-0.01 [-0.15, 0.12]
Item18	6.35	6.45	t(808.44) = -1.24; p =0.217	-0.09 [-0.22, 0.05]
Item19	6.35	6.42	t(804.62) = -0.85; p =0.398	-0.06 [-0.2, 0.08]
Item20	5.76	5.79	t(810.04) = -0.31; p =0.759	-0.02 [-0.16, 0.12]

Table S15.

Criteria for evaluation of replications by LeBel et al. (2018). A classification of relative methodological similarity of a replication study to an original study. “Same” (“different”) indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. “Everything controllable” indicates design facets over which a researcher has control. Procedural details involve minor experimental particulars (e.g., task instruction wording, font, font size, etc.).

Target similarity	Highly similar				Highly dissimilar
Category	Direct replication				Conceptual replication
Design facet	**Exact replication**	**Very close replication**	**Close replication**	**Far replication**	**Very far replication**
Effect/ Hypothesis	Same/similar	Same/similar	Same/similar	Same/similar	Same/similar
IV operationalization	Same/similar	Same/similar	Same/similar	Different	Different
DV operationalization	Same/similar	Same/similar	Same/similar	Different	Different
IV stimuli	Same/similar	Same/similar	Different	Different	
DV stimuli	Same/similar	Same/similar	Different		
Procedural details	Same/similar	Different			
Physical setting	Same/similar	Different			
Contextual variables	Different				

Table S16

Classification of the two replications based on LeBel et al.'s (2017) taxonomy.

Design facet	Study 1 Hong Kong replication	Study 2 MTurk replication
IV operationalization	same	same
DV operationalization	same	same
IV stimuli	same	same
DV stimuli	same	same
Procedural details	similar	similar
Physical setting	different	different
Contextual variables	different	different
Replication classification	Very close replication	Very close replication

Table S17.

The results of mixed-effects regression analysis with intentionality ratings as the DV.

	Original Study				**Replication Study 1**				**Replication Study 2**			
Fixed effect predictors	***B***	***S.E.***	***CI***	***p***	***B***	***S.E.***	***CI***	***p***	***B***	***S.E.***	***CI***	***p***
(Intercept)	4.11	0.45	3.22 – 4.99	**<0.001**	3.79	0.34	3.11 – 4.46	**<0.001**	4.40	0.44	3.55 – 5.26	**<0.001**
Perspective (observer vs. actor)	0.09	0.14	-0.19 – 0.37	0.529	0.33	0.17	-0.00 – 0.66	0.053	0.13	0.04	0.04 – 0.22	**0.003**
Definition (with vs. without)	-0.10	0.13	-0.32 – 0.20	0.634	0.09	0.17	-0.24 – 0.42	0.591	-0.06	0.04	-0.15 – 0.02	0.152
Random Effects												
σ^2	2.35				2.99				2.5			
$\tau_{00\ ID}$	0.29				0.18				0.27			
$\tau_{00\ Item\ ID}$	3.81				1.92				3.8			
ICC	0.64				0.41				0.62			
$N_{Item\ ID}$	20				20				20			
N_{ID}	104				46				817			
Observations	2079				919				16340			
Marginal R^2 / conditional R^2	0.000 / 0.636				0.006 / 0.415				0.001 / 0.620			

Note: N (original study) = 104; N (Study 1) = 46; N (Study 2) = 817.

We regressed intentionality ratings (DV) on perspective (observer vs. actor) and definition (with definition vs. without definition), adjusting for covariates at Level 2 (participant's ID and item identifier). The results do not show support for presence or absence of a definition as a fixed effect predictor of intentionality ratings. We found support for the type of perspective as a fixed effect predictor in Study 2, but not in the original study and Study 1.

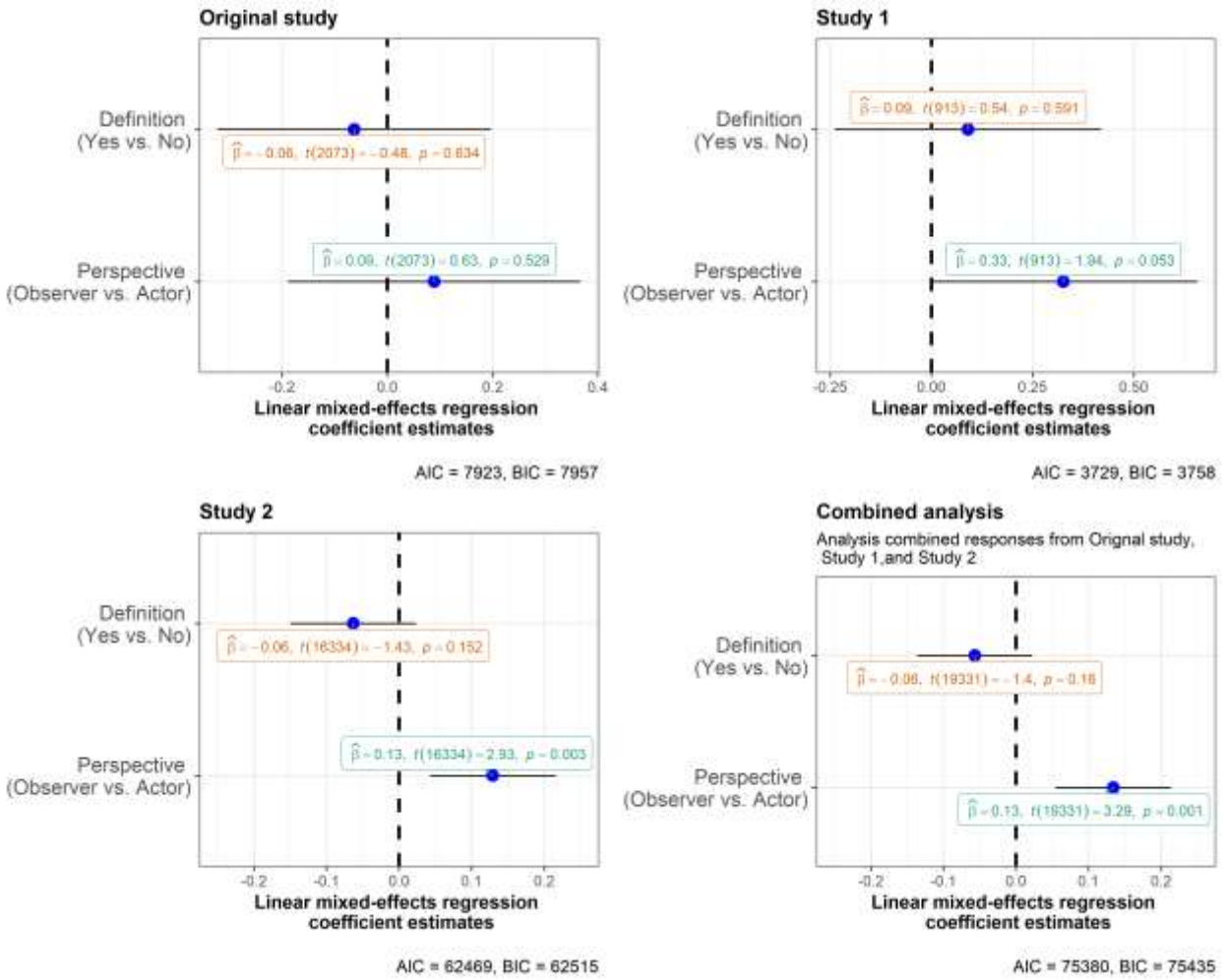


Figure S2. Regression coefficients from linear mixed-effects regressions analyses. The plots were created using *ggstatsplot* R package (Patil, 2021).

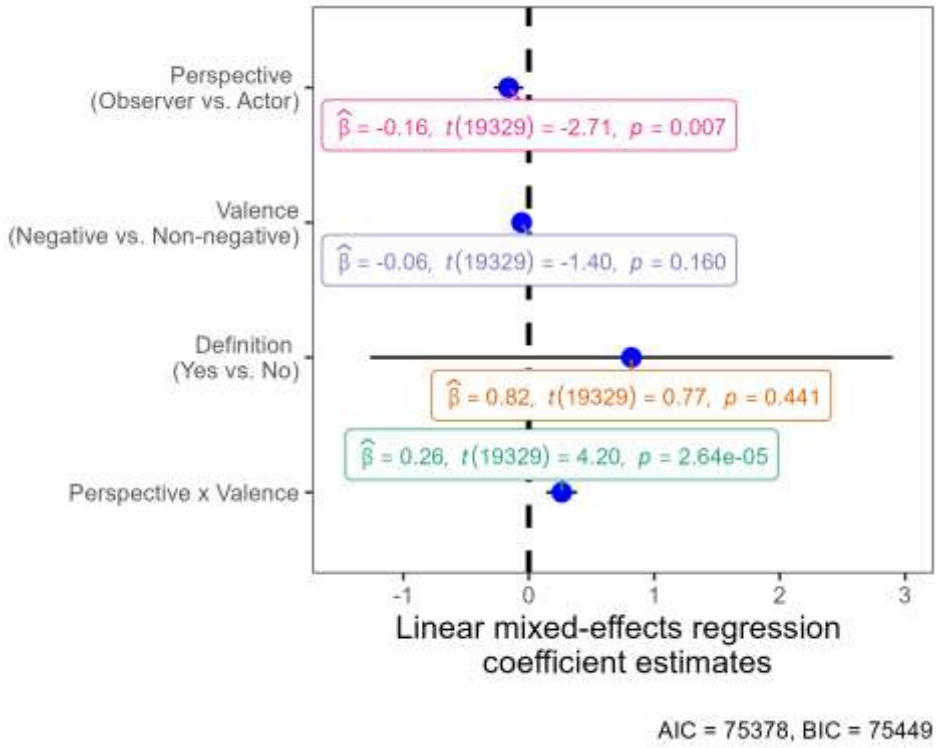


Figure S3. Regression coefficients from linear mixed-effects regressions analyses (See Model 2 of Table 3 in the Manuscript document). The plots were created using *ggstatsplot* R package (Patil, 2021).

Table S18.

The results of mixed-effects regression analysis with intentionality ratings as the DV. For this analysis we removed responses to ‘item 8’, ‘item 12’, ‘item 18’, and ‘item 19’.

	DV: Intention Rating			
Predictors	***Estimates***	***S.E.***	***CI***	***p***
(Intercept)	3.93	0.47	3.00 – 4.85	**<0.001**
Perspective (observer vs. actor)	0.14	0.05	0.05 – 0.24	**0.004**
Definition (with vs. without)	-0.10	0.05	-0.20 – -0.01	**0.037**
Random Effects				
σ^2			2.73	
$\tau_{00\ ID}$			0.33	
$\tau_{00\ ItemID}$			3.52	
ICC			0.58	
$N_{\ ItemID}$			16	
$N_{\ ID}$			817	
Observations			13072	
Marginal R^2 / conditional R^2			0.001 / 0.585	

References

Ames, D. R., Knowles, E. D., Morris, M. W., Kalish, C. W., Rosati, A. D., & Gopnik, A. (2001). The social folk theorist: Insights from social and cultural psychology. In B. F.,Malle, L. J., Moses, & D. A., Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 307-328). Cambridge, MA: MIT Press.

Bagby, R. M., Parker, J. D., & Taylor, G. J. (1994). The twenty-item Toronto Alexithymia Scale—I. Item selection and cross-validation of the factor structure. *Journal of Psychosomatic Research*, *38*, 23-32. Doi: 10.1016/0022-3999(94)90005-1

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, 10, 535-549. Doi: 10.1111/spc3.12267

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, 8, 875-881. Doi: 10.1177/1948550617693058

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://www.jstatsoft.org/v36/i03/.

jamovi project (2018). jamovi (Version 0.8) [Computer Software]. Retrieved from https://www.jamovi.org

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, 113, 254–261. Doi: 10.1037/pspi0000106

Malle, B. F., & Knobe, J. (1997). The folk concept of intentionality. *Journal of Experimental Social Psychology*, *33*, 101-121. Doi: 10.1006/jesp.1996.1314

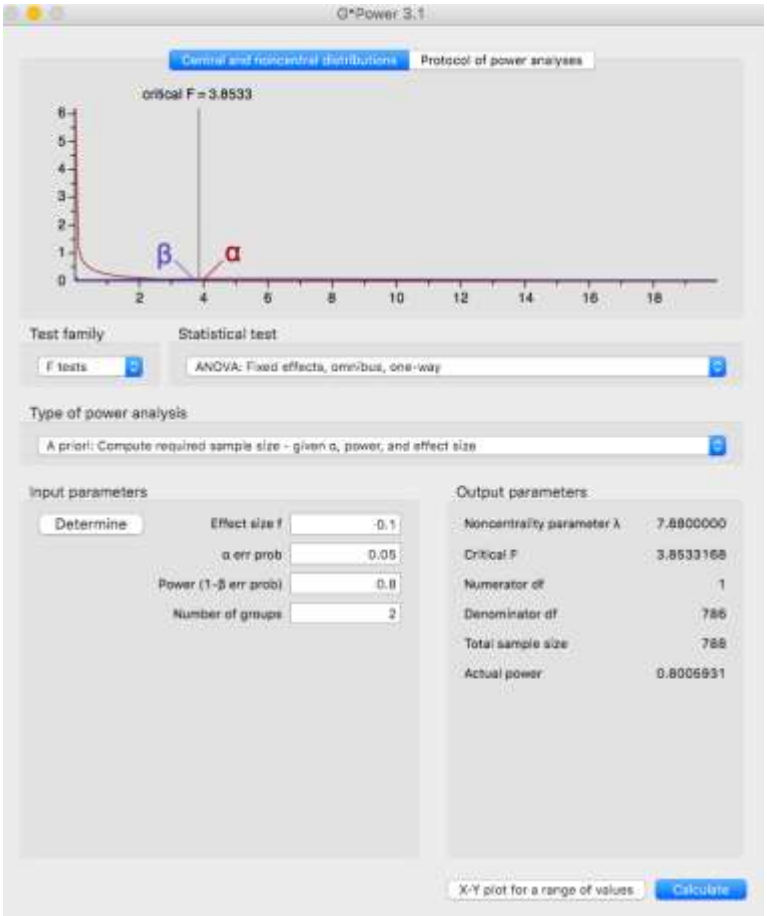
Ohtsubo, Y. (2007). Perceived intentionality intensifies blameworthiness of negative behaviors: Blame-praise asymmetry in intensification effect 1. *Japanese Psychological Research*, *49*, 100-110. Doi: 10.1111/j.1468-5884.2007.00337.x

Appendix A

Effect Sizes Calculations and Power Analyses

Malle and Knobe (1997) Study 1 hypothesized and empirically found that perspective (actor vs. observer) and provision of definition (with vs. without) did not have significant main effects on intentionality ratings. Accordingly, we expected weak effects of perspective and provision of definition on intentionality rating. Cohen (1988) considered effect sizes between -.20 and .20 as weak. Thus, we used Cohen's d = .20, which is equivalent to Cohen's f = .10, to estimate the required sample size for power = 80% and α = .05.

Software: G*Power Version 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007).



F tests - ANOVA: Fixed effects, omnibus, one-way

Analysis: A priori: Compute required sample size

Input:	Effect size f	=	0.1
	α err prob	=	0.05
	Power (1-β err prob)	=	0.8
	Number of groups	=	2
Output:	Noncentrality parameter λ	=	7.8800000
	Critical F	=	3.8533168
	Numerator df	=	1
	Denominator df	=	786
	Total sample size	=	788
	Actual power	=	0.8005931