Perspectives on Psychological Science
2016, Vol. 11(1) 158–171
© The Author(s) 2015
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1745691615605826
pps.sagepub.com

Registered Replication Report: Hart & Albarracín (2011)

A. Eerland*, A. M. Sherrill*, J. P. Magliano*, R. A. Zwaan*, J. D. Arnal, P. Aucoin, S. A. Berger, A. R. Birt, N. Capezza, M. Carlucci, C. Crocker, T. R. Ferretti, M. R. Kibbe, M. M. Knepp, C. A. Kurby, J. M. Melcher, S. W. Michael, C. Poirier, J. M. Prenoveau
*Proposing authors

Protocol vetted by: William Hart
Protocol edited by: Alex O. Holcombe

Multilab direct replication of: Study 3 from Hart, W., & Albarracín, D. (2011). Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science, 22*, 261–266.

Data and registered protocols: https://osf.io/d3mw4/

Citation: Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., . . . Prenoveau, J. M. (2016). Registered replication report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*, 158–171.

Abstract

Language can be viewed as a complex set of cues that shape people's mental representations of situations. For example, people think of behavior described using imperfective aspect (i.e., what a person *was doing*) as a dynamic, unfolding sequence of actions, whereas the same behavior described using perfective aspect (i.e., what a person *did*) is perceived as a completed whole. A recent study found that aspect can also influence how we think about a person's intentions (Hart & Albarracín, 2011). Participants judged actions described in imperfective as being more intentional (d between 0.67 and 0.77) and they imagined these actions in more detail (d = 0.73). The fact that this finding has implications for legal decision making, coupled with the absence of other direct replication attempts, motivated this registered replication report (RRR). Multiple laboratories carried out 12 direct replication studies, including one MTurk study. A meta-analysis of these studies provides a precise estimate of the size of this effect free from publication bias. This RRR did not find that grammatical aspect affects intentionality (d between 0 and −0.24) or imagery (d = −0.08). We discuss possible explanations for the discrepancy between these results and those of the original study.

Keywords
grammatical aspect, behavioral intentionality, attribution, legal psychology, replication

Address correspondence to:
Anita Eerland, Trans 10, 3512JK, Utrecht, Netherlands
E-mail: a.eerland@uu.nl

People use language to convey ideas about situations in the real world or in some hypothetical world. Language can be viewed as a complex set of cues that help shape how people understand and represent actions and events in their world (Johnson-Laird, 1983; Morrow, Greenspan, & Bower, 1987; Van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998). Grammatical aspect is one such cue. Behavior described using imperfective aspect (i.e., what a person *was doing*) is perceived as a dynamic, unfolding sequence of actions, whereas the same behavior described using perfective aspect (i.e., what a person *did*) is perceived as a completed whole (Madden & Zwaan, 2003; Magliano & Schleich, 2000; Mozuraitis, Chambers, & Daneman, 2013). Actions described in imperfective aspect are perceived as more vivid and perceptually engaging than are those in perfective aspect. Perhaps this enhancement of detailed processing (i.e., richer encoding) causes actions described in imperfective aspect to also be more memorable than other aspectual forms (Carreiras, Carriedo, Alonso, & Fernandez, 1997; Magliano & Schleich, 2000).

The results of a recent study suggest that grammatical aspect not only influences our understanding of and memory for described situations, but also whether we think of an action as being intentional (Hart & Albarracín, 2011). The first two experiments showed that grammatical aspect influences our perceptions of intentionality for mundane behavior. For example, whenever participants read sentences describing the actions of a person in imperfective aspect (e.g., Keith was preparing dinner for some friends) compared with those describing actions in perfective aspect (e.g., Keith prepared dinner for some friends), they were more likely to complete word stems with intention-relevant words. Experiment 3, which is the focus of this RRR, showed that grammatical aspect influences perceived intentionality for both mundane and criminal behavior. In the experiment, participants read a vignette (see Appendix A) about a man being shot by another man after the two had argued about a dice game. The actions of the perpetrator were either described in imperfective aspect (i.e., pulling out a gun, pointing it at the other man, and shooting the gun) or perfective aspect (i.e., pulled out a gun, pointed it at the other man, and shot the gun). Participants rated the perpetrator's harmful intent higher when the actions were described in imperfective aspect rather than in perfective aspect. Mediation analyses indicated that imperfective aspect resulted in higher intentionality ratings because it promoted more detailed processing of the described criminal acts.

The finding that subtle shifts in aspect can change how people interpret intentions has implications for explaining, predicting, and morally judging the behavior of others (Young & Waytz, 1993), such as in legal decision making (e.g., a prosecutor could use imperfective aspect to imply greater intentionality by the accused suspect). Moreover, whereas other known effects of grammatical aspect on situation models tend to be rather small (e.g., Ferretti, Kutas, and McRae (2007) found Cohen's $d = 0.16$), this study found large effects of aspect on mundane behavior (d between 0.99 and 1.03) and moderate to large effects on criminal behavior (d between 0.67 and 0.77).[1] The size of these effects suggests that differences in the use of grammatical aspect could have far-reaching practical consequences.

To date, no independent replication exists of this finding, supporting the case for a registered replication report (RRR). RRR projects are designed to arrive at a precise estimate of the magnitude of a previously reported effect by meta-analyzing a set of replications of the original study. Multiple laboratories independently conduct a direct replication of the same study by following a preregistered protocol. By following this procedure, the specific goal of this RRR was to provide an accurate estimate of the effect of grammatical aspect on perceived intentionality.

Protocol Development to Compare Past and Present Studies

For a direct replication of the original grammatical aspect study, Eerland, Sherrill, Magliano, and Zwaan developed the protocol in consultation with the original study's first author, William Hart. The protocol was designed to follow the original study's methodology as closely as possible, with one exception described in *Materials* below. After finalizing the protocol, *Perspectives on Psychological Science* publicly announced a call for laboratories interested in participating in this replication project on March 6, 2014. A deadline for applications to participate was set for April 10, 2014, and by that time 11 labs joined this project. All labs conducted an independent replication and preregistered their plan for implementing the protocol. Each implementation plan was checked by the editor for agreement with the protocol before the start of data collection.

Laboratories in the United States, Canada, and the Netherlands participated. The researchers from the Netherlands ran a large-scale online experiment recruiting U.S. participants from Amazon Mechanical Turk (MTurk; http://www.mturk.com) so that the materials of the original study did not require any translation. This online experiment differed slightly from the lab-based studies (see *Online Version of the Protocol*). Most labs included experts on language comprehension, memory, and/or forensic psychology, and all are coauthors on this manuscript. Some labs lacked domain-specific experience, but all were experienced in conducting psychology experiments.

Protocol Requirements

Following best practice, we report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons, Nelson, & Simonsohn, 2012). Each participating laboratory created a collection of webpages on the Open Science Framework (linked from a master page https://osf.io/d3mw4/) on which they posted their implementation plan and results. The constraints of the research protocol are described below.

Participants

The protocol specified a minimum sample size of 30 participants per condition, and we encouraged labs to include as many participants as possible. As in the original study, participants were drawn from an undergraduate subject population, most participants within each sample had to be between 18 and 25 years of age, and no more than 50% of the sample could be male. Given that the methods are language sensitive, all participants had to report English as their first and primary language.

Testing location

Participants were tested individually and in small groups of up to 6 participants (6 labs) or individually only (5 labs). When tested in groups, participants were prevented from seeing, hearing, or communicating with other participants during the experiment. One lab recruited from a university community-specific subject pool but tested them online. The proposing authors conducted their version of the study using MTurk (see *Online Version of the Protocol*).

Experimenters

Any trained research assistant, postdoctoral researcher, or faculty member could serve as the experimenter as long as they had experience interacting with and testing participants and had experience using the software that was chosen to program the experiment. Experimenters were aware that participants were randomly assigned to one of two conditions, but they were blind to each participant's condition assignment.

Materials

The study used the vignettes and questions from the original article. After consulting with the editor and Hart, we made one change to the imperfective-aspect version of the vignette. We changed "Westmoreland was pulling out his gun and was pointing it at Darryl McElroy" into "Westmoreland pulled out his gun and was pointing it at Darryl McElroy" because the original wording suggested that Westmoreland was simultaneously pulling out and pointing the gun rather than performing one action and then the other. Hart provided additional questions that participants in the original study had completed but that were not reported in the original article. He also supplied the instructions and the order of the questions of the original study. We added one question at the end of the experiment to investigate the following possible explanation for the results that was raised in the original article: "compared with the perfective aspect, the imperfective aspect may suggest a longer behavior duration, which may in turn suggest greater persistence and intent" (Hart & Albarracín, 2011, p. 256).

The experiment, including the vignette and all questions, was programmed and presented in the Qualtrics survey research suite (http://www.qualtrics.com). The scripts used to run the experiment are available from https://osf.io/d3mw4/.

Data collection

Each participant viewed the instructions, vignette, and questions on their own computer screen. The questions appeared in a fixed and predetermined order and were answered using the computer keyboard and mouse. Participants did not know they were participating in a study on the effects of grammatical aspect on intentionality, and they were not informed of the hypotheses until after they had completed the study. Participants were randomly assigned to conditions (perfective vs. imperfective) with the constraint that approximately equal numbers of participants were assigned to each condition.

Procedure

Participants read the following instructions: "Below is a brief case report that involves James Westmoreland. Please imagine that you are a judge interested in making the best sentencing decision for Westmoreland based on the following case report." Then they were presented with either the perfective aspect or the imperfective aspect version of the vignette. After reading this vignette, participants were asked questions about how they would sentence Westmoreland and their opinion of his mental state during the crime. Participants indicated on an 11-point scale to what extent Westmoreland knowingly/intentionally/deliberately caused harm to McElroy (−5 = *unknowingly/unintentionally/undeliberately*, 5 = *knowingly/intentionally/deliberately*). The mean score of these three questions was used as measure of criminal intentionality. Further, participants indicated whether they thought Westmoreland should serve time in prison

(*yes/no*) and, if so, how many years in prison he should serve (any whole number between 1–70). Then all participants saw the following question on the screen: "In your opinion, how should Westmoreland be punished for his crime? Please indicate one of the following punishment options: 1) Probation with no prison term, 2) 5 years with the opportunity for parole, 3) 10 years with the opportunity for parole, 4) 15 years with the opportunity for parole, 5) 20 years with the opportunity for parole, 6) 25 years with the opportunity for parole, 7) Life imprisonment with the opportunity for parole, 8) Life imprisonment with NO opportunity for parole, 9) Death penalty sentence."

Next, participants read the following instructions: "For the next questions, we'd like to ask you about your experience reading the case report. There is no right or wrong answer to any of the questions so you should just provide your genuine reaction to each question. Here again is the case report you read." The same criminal report was shown on screen and participants indicated to what extent the case report made it easy or difficult to imagine the crime unfolding/Westmoreland's concrete behaviors/Westmoreland's physical movements/the details of the crime on a 7-point scale (1 = *very easy*, 7 = *very difficult*). The mean score of these four items resulted in a measure of detailed processing.

In addition, participants used a 5-point scale to indicate the extent to which they agreed or disagreed with 10 statements that were taken from the Mind Attribution Scale (Kozak, Marsh, & Wegner, 2006; 1 = *strongly disagree*, 5 = *strongly agree*). The mean score of the first three items was used as measure of intention attribution.

Then, we asked participants to indicate how many gunshots Westmoreland fired. This question was added to address the possible explanation for the results noted above (see *Materials*). Finally, we asked participants to describe what they thought the study was about and to report their gender, age, and native language.

Data collection stopping rules and exclusions

Each lab pre-registered (a) their stopping rules for data collection, (b) how they would ensure that they met the demographic requirements of the protocol, (c) how they would assess the first and primary language of participants, (d) how participants would be assigned to conditions, and (e) how additional data would be collected if participants had to be excluded. The editor reviewed these procedures to verify that they ensured pseudo-random assignment to conditions and that each lab would be able to meet the minimum required sample size after any exclusions necessitated by the protocol requirements.

The following were acceptable reasons to exclude participants: reporting any other language than English as their native language, being younger than 18 years of age, not completing the study, not following instructions (e.g., advancing to the next screen without reading the instructions), or experimenter error. Participants older than 25 years of age could be included as long as most participants in the sample were between 18 and 25 years of age. All decisions about the criteria to use for excluding data were made before data collection began. The raw data files (see OSF pages) include the data that were excluded from analysis and also report the reason for exclusion.

Online experiment

A large-scale online replication study was conducted by the lead authors on the RRR. Participants for this online study were recruited from MTurk and were limited to native English speakers who resided in the United States and had a Human Intelligence Task (HIT) approval rate of >95%. Given these less restrictive inclusion criteria, participants in this online study are likely to differ from those in the original study and the other in-lab studies[2] with respect to age, education, and gender distribution. Also in contrast to the in-lab studies, this online study involved no contact between participants and the experimenter. Given the difference in setting, we added a question to this study asking about the environment in which participants completed the experiment. In all other ways, the materials and procedure for the MTurk study were the same as those for the in-lab replication studies.

The results of the MTurk study were not included in the meta-analysis of the in-lab-studies. However, they are reported along with the other lab results in Table 1 and in all figures.

Results

Lab demographics and results

Some demographics and the results of each participating lab as well as for the original study are provided in Table 1. This table also includes the number of participants tested in each condition, the number of excluded participants (and the reason for exclusion), and the mean and standard deviation for all three outcome measures in each condition. Demographics for all participating labs can be found in Appendix B.

Data analyses: Original and RRR

The original article reported three separate ANOVAs comparing the imperfective aspect condition to the perfective aspect condition for criminal intentionality,

Table 1. Sample Sizes and Data for the Original Study, the Online MTurk Replication Study, and All In-Lab Replication Studies

		Perfective aspect condition								Imperfective aspect condition							
Lab	Country of participants	Total *N*	Excluded age	Excluded native language	Excluded other	Total included	Criminal intentionality *M* (*SD*)	Intention attribution *M* (*SD*)	Detailed processing *M* (*SD*)	Total *N*	Excluded age	Excluded native language	Excluded other	Total included	Criminal intentionality *M* (*SD*)	Intention attribution *M* (*SD*)	Detailed processing *M* (*SD*)
Original study, W. Hart and D. Albarracín (2011)	USA	24	0	0	0	24	3.69 (1.94)	4.06 (0.79)	4.48 (1.34)	24	0	0	0	24	4.89 (1.06)	4.61 (0.84)	5.40 (1.16)
Online study (MTurk), A. Eerland, A. M. Sherrill, J.P. Magliano, and R.A. Zwaan	USA	131	0	1	0	130	4.54 (1.03)	4.13 (0.58)	5.90 (0.90)	143	0	7	6	130	4.21 (1.36)	4.15 (0.59)	5.77 (0.97)
J.D. Arnal	USA	37	0	2	0	35	3.82 (1.37)	4.03 (0.57)	4.88 (1.13)	36	0	4	0	32	3.47 (1.61)	3.88 (0.54)	4.87 (1.12)
S.A. Berger	USA	82	0	45	2	35	4.16 (1.00)	3.83 (0.54)	5.34 (0.97)	80	0	37	3	40	3.18 (2.12)	3.68 (0.53)	4.89 (1.16)
A.R. Birt and P. Aucoin	Canada	36	1	2	0	33	4.13 (1.18)	3.83 (0.53)	4.88 (0.93)	34	0	1	0	33	3.75 (1.27)	3.91 (0.63)	4.48 (1.10)
A. Eerland, A.M. Sherrill, J.P. Magliano, and R.A. Zwaan	USA	59	4	5	0	50	3.63 (1.72)	3.76 (0.56)	5.13 (1.05)	67	4	13	0	50	3.79 (1.61)	3.75 (0.45)	5.12 (1.24)
T.R. Ferretti	Canada	42	0	0	0	42	3.94 (1.09)	3.79 (0.54)	4.94 (1.29)	44	0	1	0	43	3.93 (0.85)	3.98 (0.66)	5.27 (0.86)
M.M. Knepp	USA	47	0	4	0	43	4.00 (1.26)	3.91 (0.66)	4.91 (1.24)	43	0	0	0	43	3.05 (1.68)	3.62 (0.74)	4.91 (1.19)
C.A. Kurby, M.R. Kibbe	USA	60	0	0	0	60	3.64 (1.50)	3.94 (0.52)	4.81 (1.05)	60	0	0	0	60	3.50 (1.83)	3.94 (0.62)	4.83 (1.28)
J.M. Melcher	USA	33	0	0	0	33	2.94 (2.04)	3.82 (0.65)	5.08 (1.37)	36	0	3	0	33	3.59 (1.76)	3.94 (0.83)	4.92 (1.31)
S.W. Michael	USA	44	0	5	0	39	3.85 (1.47)	4.01 (0.56)	5.07 (1.13)	45	0	1	0	44	3.44 (1.51)	4.14 (0.58)	4.99 (1.16)
C. Poirier, N. Capezza, and C. Crocker	USA	40	0	0	0	40	3.53 (1.64)	3.78 (0.62)	5.09 (0.99)	40	0	0	1	39	3.85 (1.33)	3.84 (0.63)	4.90 (1.08)
J.M. Prenoveau and M. Carlucci	USA	74	2	3	19	50	3.57 (1.64)	3.79 (0.58)	4.79 (1.20)	53	1	0	2	50	3.19 (1.81)	3.82 (0.60)	4.77 (1.32)

detailed processing (i.e., processing of the criminal act described in the vignette), and intention attribution. Additional analyses explored whether detailed processing mediated the relation between aspect and criminal intentionality and the relation between aspect and intention attribution. For this RRR, each lab performed three independent samples *t* tests comparing the imperfective and perfective aspect conditions for the same three outcome measures and conducted two mediation analyses with bootstrapping (10,000 samples) in SPSS using the PROCESS plug-in (created by Andrew Hayes; http://www.afhayes.com/introduction-to-mediation-moderation-and-conditional-process-analysis.html). The detailed results from each study are reported on each lab's OSF page (see Appendix B for URLs). As in the original study, we used Cohen's *d* as our measure of effect size.

Effect size measurements

The results for all three measures of interest (i.e., criminal intentionality, detailed processing, and intention attribution) are each displayed in a forest plot showing the means and standard deviations in both conditions for each lab, the effect size measured by each lab (with 95% confidence intervals), and the meta-analytic effect size estimate in a random effects model (see Figs. 1–3). The top row in each plot shows the original effect reported by Hart and Albarracín, and the row below that shows the effect found in the online MTurk variant of the study. Those results are not included in the meta-analytic effect size estimate reported at the bottom of each figure—only the lab-based replications of the original study are included in the meta-analyses.

Figure 1 shows a small effect in the opposite direction of that reported in the original study for criminal intentionality, but this is not statistically significant—the 95% confidence interval includes zero. In the original study, intentionality ratings were 1.2 points higher in the imperfective condition than in the perfective condition, whereas we found an average difference of 0.24 points in the opposite direction (95% confidence interval: −0.49 to 0.03). Most labs observed similarly small differences, and 8 of the 11 lab studies as well as the MTurk study found (nonsignificant) effects that were numerically in the opposite

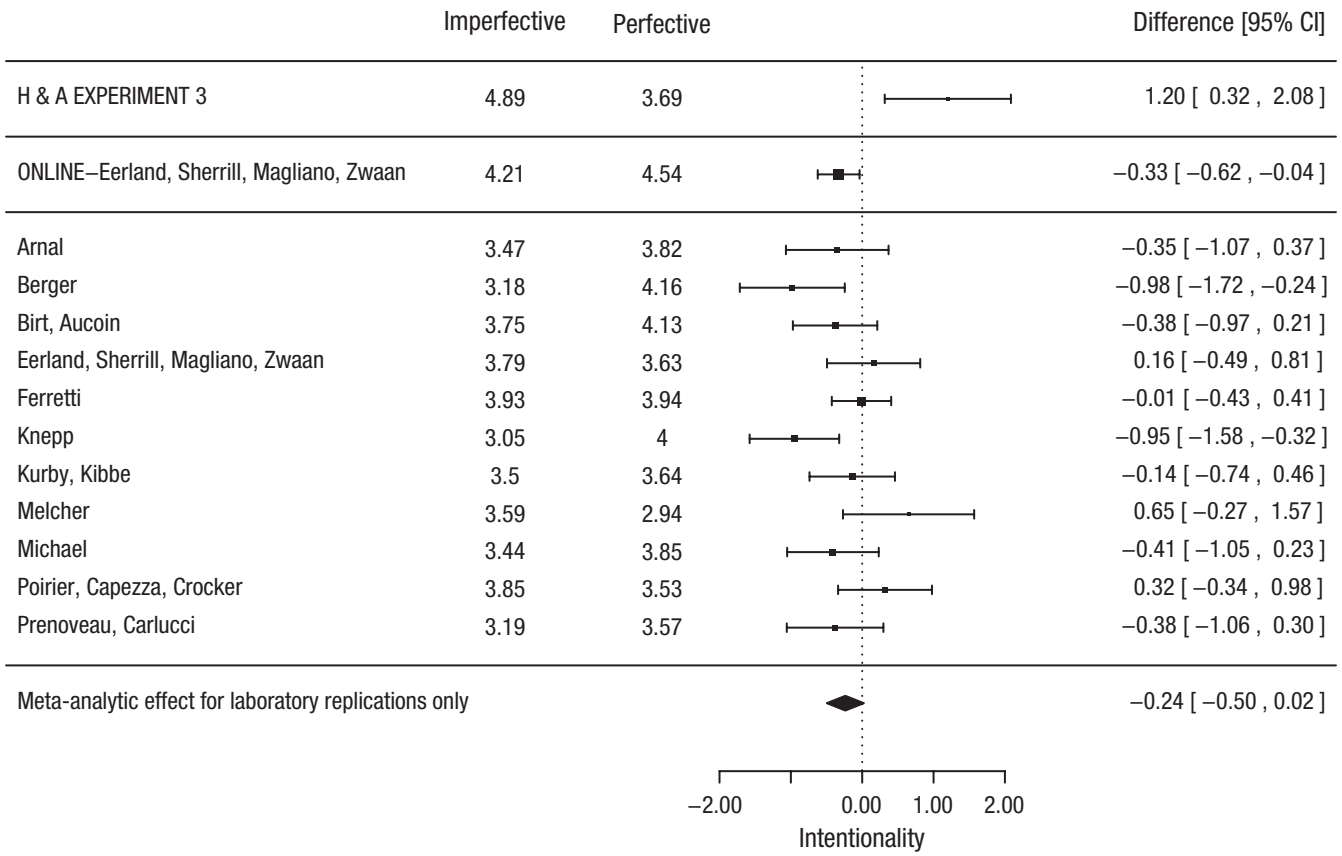

Fig. 1. Forest plot of the effect of grammatical aspect on criminal intentionality, with negative effects indicating lower scores for participants in the imperfective aspect condition than the perfective aspect condition (Imperfective – Perfective). The data are listed in alphabetical order by the name of the first author from each replicating team. For each team, the figure shows the mean criminal intentionality score for the imperfective and the perfective aspect condition and a forest plot of the raw mean difference score (bigger effect size markers reflect bigger samples). The Difference column provides the values used in the forest plot.

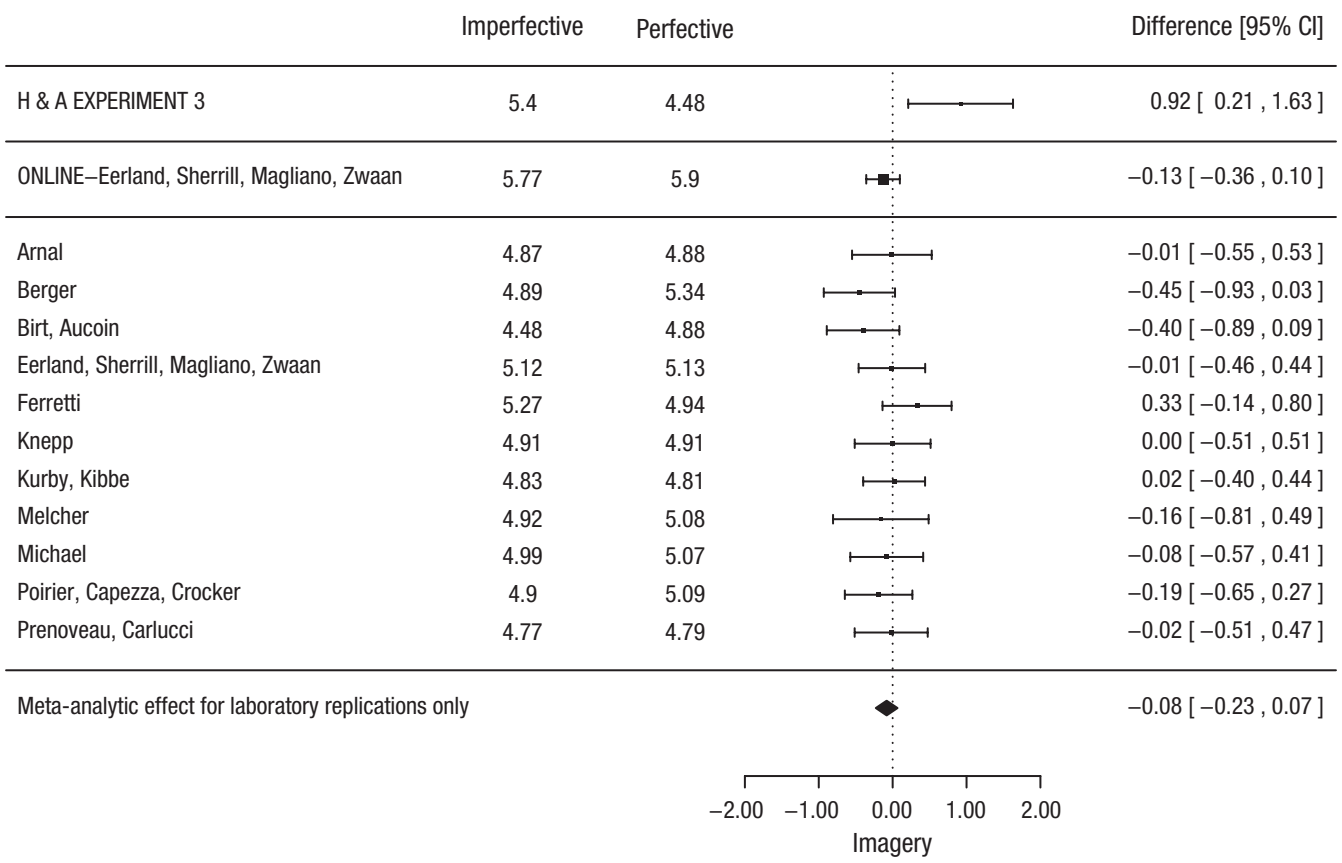

Fig. 2. Forest plot of the effect of grammatical aspect on detailed processing, with negative effects indicating lower scores for participants in the imperfective aspect condition than the perfective aspect condition (Imperfective – Perfective). The data are listed in alphabetical order by the name of the first author from each replicating team. For each team, the figure shows the mean detailed processing score for the imperfective and the perfective aspect condition and a forest plot of the raw mean difference score (bigger effect size markers reflect bigger samples). The Difference column provides the values used in the forest plot.

direction of the original finding. All the replication effect size estimates, including the online MTurk study, fell between –0.98 and 0.65. The differences in the estimated effect size among the studies (i.e., heterogeneity) were larger than what would be expected by chance ($\tau = 0.29$, $I^2 = 44.94\%$, $H^2 = 1.82$, $Q_{10} = 18.626$, $p = .045$).[3] Typically, this indicates the instability of an effect (i.e., variation in true effect sizes being measured by different studies) and/or the influence of a moderator. Although all lab studies were run using comparable participants and conditions, we cannot rule out that the heterogeneity is due to a (not yet identified) moderator, such as regional differences in "conservativeness" (Hart & Albarracín, 2011).

The meta-analysis displayed in Figure 2 also shows a small effect in the opposite direction to that reported in the original study for detailed processing. In the original study, actions in imperfective aspect were easier to imagine than actions in perfective aspect, with ratings of detailed processing that were 0.92 points higher for actions described in imperfective aspect than for actions described in perfective aspect. Our meta-analysis found a difference of 0.08 (95% confidence interval: –0.23 to 0.07) favoring the perfective aspect condition. All of the replication effect size estimates, including the online MTurk study, fell between –0.45 and 0.33. The differences in the estimated effect size among the studies (i.e., heterogeneity) were consistent with what would be expected by chance ($\tau = 0$, $I^2 = 0\%$, $H^2 = 1.00$, $Q_{10} = 7.656$, $p = .662$).

The meta-analysis displayed in Figure 3 shows an effect of grammatical aspect on intention attribution that is close to zero. The original study found that intention attribution ratings were 0.55 points higher in the imperfective aspect condition than in the perfective aspect condition. Our meta-analysis yielded a difference of 0.001 (95% confidence interval: –0.08 to 0.08) between conditions. Most individual studies, including the MTurk study, showed a small effect in the same direction as the original finding. All of the replication effect size estimates, including the online MTurk study, fell between –0.29 and 0.19. The differences in the estimated effect size among the studies (i.e., heterogeneity) were consistent with what would be expected by chance ($\tau = 0$, $I^2 = 0.02\%$, $H^2 = 1.00$, $Q_{10} = 10.554$, $p = .393$).

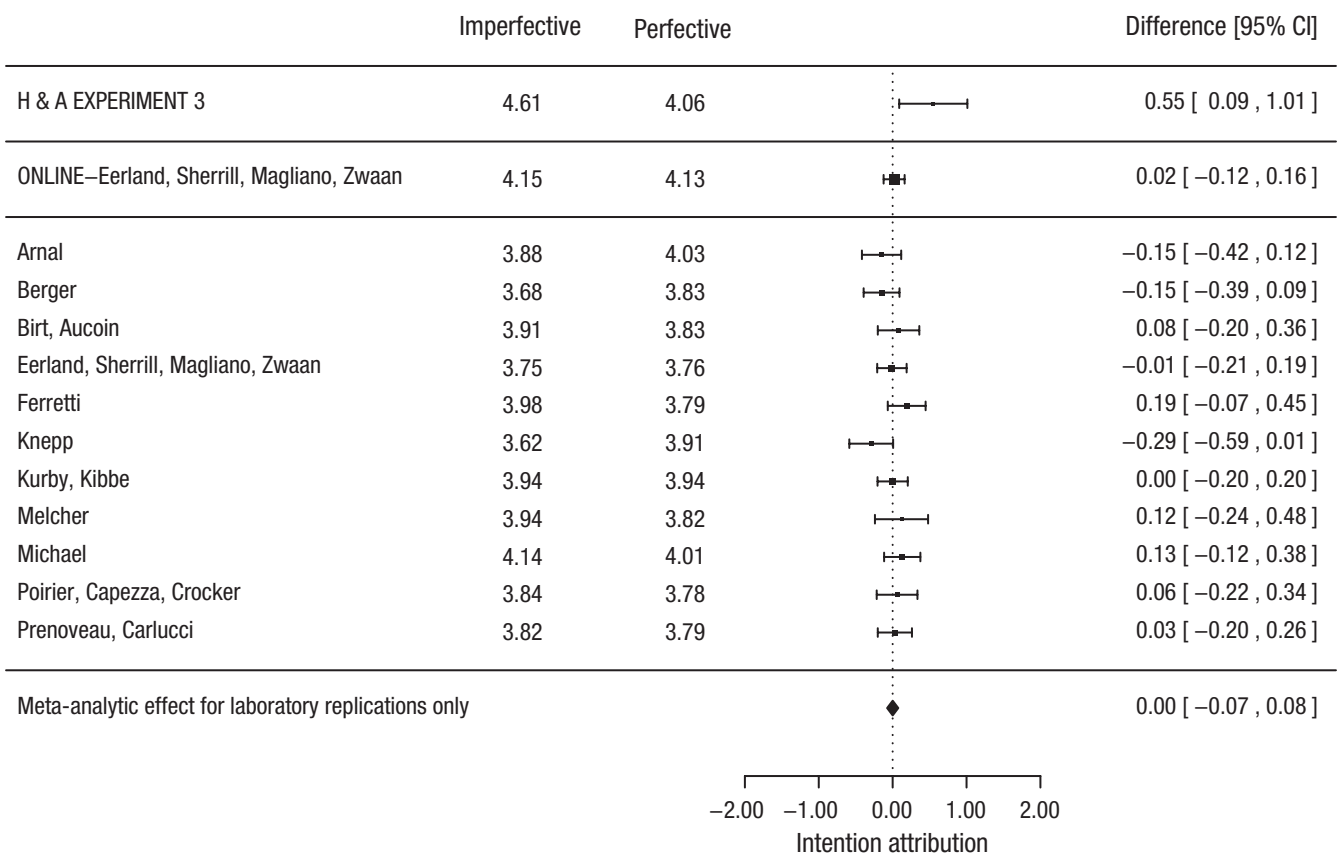

Fig. 3. Forest plot of the effect of grammatical aspect on intention attribution, with negative effects indicating lower scores for participants in the imperfective aspect condition than the perfective aspect condition (Imperfective – Perfective). The data are listed in alphabetical order by the name of the first author from each replicating team. For each team, the figure shows the mean intention attribution score for the imperfective and the perfective aspect condition and a forest plot of the raw mean difference score (bigger effect size markers reflect bigger samples). The Difference column provides the values used in the forest plot.

Taken together, our studies did not provide evidence that describing actions in imperfective aspect resulted in greater perceived intentionality (i.e., criminal intentionality and intention attribution) or more detailed processing of those actions. The overall pattern of results was consistent across studies, including the MTurk study. Indeed, the results of the MTurk study closely matched the outcomes of the meta-analyses.

Mediation effects

As in the original study, we performed mediation analyses to investigate the effect of detailed processing on the relation between grammatical aspect and criminal intentionality and grammatical aspect and intention attribution. The effect found in the online MTurk variant of the study was not included in this meta-mediation analysis, nor is the replication study by Knepp. The Knepp study found identical means for both conditions on the measure of detailed processing, thus precluding mediation analyses.

Note that some argue against performing mediation analysis in the absence of clear relations between (a) the independent variable and the dependent variable (i.e., grammatical aspect and criminal intentionality/intention attribution), (b) the independent variable and the mediator (i.e., grammatical aspect and detailed processing), and (c) the mediator and the dependent variable (i.e., detailed processing and criminal intentionality/intention attribution; Baron & Kenny, 1986). All of these relations were present in the original study but not in our meta-analyses. Although more recent work shows that mediation might still occur when there is no significant relation between the independent and the dependent variable, this typically happens when analyses are underpowered (Hayes, 2009; Shrout & Bolger, 2002). This is not the case for our meta-analyses. However, the independent samples *t* tests used for the primary analyses might not be sensitive enough to pick up on any mediation effect, so meta-mediation analyses could be informative. More important, the mediation analyses were part of the preregistered analysis plan.

We performed meta-mediation analyses using Exploratory Software for Confidence Intervals (ESCI; Cumming, 2012). We used Cohen's *d* effect sizes as input for our analyses. (Hart provided the R^2 values of the indirect effects found in the original study to enable us to calculate Cohen's *d*.)

Although the original study found that detailed processing partly explained the relation between grammatical aspect and criminal intentionality ($b = 0.50$, $SE = 0.28$, Cohen's $d = 0.64$), our meta-mediation analysis found no evidence that detailed processing mediates this relation (Cohen's $d = 0.02$, 95% bias-corrected bootstrap confidence interval: −0.12 to 0.15). All of the replication effect size estimates, including the online MTurk study, fell between −0.15 and 0.31, and the differences in the estimated effect size among the studies (i.e., heterogeneity) were consistent with what would be expected by chance ($T^4 = 0$, $I^2 = 0\%$, $H^2 = 1.00$, $Q_9 = 3.303$, $p = .951$).

The second meta-mediation analysis produced similar results. In the original study detailed processing partly explained the relation between grammatical aspect and intention attribution ($b = 0.24$, $SE = 0.12$, Cohen's $d = 0.57$). In the meta-mediation, though, detailed processing did not mediate that relation (Cohen's $d < 0.001$, 95% bias-corrected bootstrap confidence interval: −0.14 to 0.14). Again, the MTurk study found comparable results. All of the replication effect size estimates, including the online MTurk study, fell between −0.11 and 0.24. The differences in the estimated effect size among the studies (i.e., heterogeneity) were consistent with what would be expected by chance ($T = 0$, $I^2 = 0\%$, $H^2 = 1.00$, $Q_9 = 1.928$, $p = .993$).

Additional direct replication

In addition to the replication studies described above, Kurby and Kibbe ran a paper-and-pencil version of the experiment to address whether the change to one sentence of the vignette might explain the different pattern of results observed in the RRR and the original study. In the Kurby and Kibbe study, participants read the original, unaltered vignette used by Hart and Albarracín (2011, see Appendix A), and they also had access to the vignette while making their ratings. This follow-up study was not preregistered and is therefore not included in the meta-analyses. A more detailed description of this study as well as the raw data are available from https://osf.io/8np9f/.

The results of this direct replication showed an effect of grammatical aspect on criminal intentionality close to zero, $t(42) = 0.154$, $p = .879$, Cohen's $d = 0.05$, 95% confidence interval: −0.87 to 0.75. The effects on grammatical aspect on detailed processing, $t(42) = 1.093$, $p = .281$, Cohen's $d = 0.33$, 95% confidence interval: −1.11 to 0.33, and intention-attribution, $t(42) = 0.921$, $p = .362$, Cohen's $d = 0.28$, 95% confidence interval: −0.51 to 0.19, were small and in the opposite direction of those observed in the original study. The mediation analyses (10,000 bootstrap samples using PROCESS) showed that detailed processing did not mediate the relation between grammatical aspect and criminal intentionality ($b = -0.07$, $SE = 0.12$, Cohen's $d = 0.05$, 95% bias-corrected bootstrap confidence interval: −0.53 to 0.05) or intention attribution ($b = 0.03$, $SE = 0.05$, Cohen's $d = 0.16$, 95% bias-corrected bootstrap confidence interval: −0.03 to 0.02).

Discussion

The results of this large-scale, multilab direct replication of Experiment 3 by Hart and Albarracín (2011) are not consistent with the original result that actions described in imperfective aspect are considered to be more intentional than actions described in perfective aspect. Most labs found effects of aspect on intentionality that were close to zero. For the criminal intentionality measure, eight labs and the MTurk study found effects that were numerically in the opposite direction of the effect reported in the original study. Four labs observed effects that were numerically in the opposite direction for the intention attribution measure.

This RRR also did not find that actions described in imperfective aspect are processed in greater detail than those described in perfective aspect. Most labs found effects that were close to zero, with eight lab studies and the MTurk study observing effects that were numerically in the opposite direction of the original study. Finally, none of the replication studies found that detailed processing mediated the relation between grammatical aspect and intentionality. What factors might account for the difference between the RRR results and those of the original study?

Random variation

One possible explanation for the differing results is that the original study observed large effects due to measurement or sampling variation. This variation makes it possible to find a large effect even when the true effect is close to zero—a so-called "false positive."

Several factors might have contributed to such variability in the original study. First, the study used a between-subjects comparison across conditions and tested the effect of aspect using just one vignette to manipulate aspect. Within-subject manipulations help to control for individual differences across conditions that could otherwise contribute to a spuriously large difference between conditions, and the use of multiple observations or vignettes provides a more stable estimate of

the effect. Studies of the influence of grammatical aspect on situation models often use multiple examples and within-participant designs, and they tend to find smaller effects (e.g., Ferretti et al., 2007, found an effect size of 0.16). With the reduced power of between-subjects designs and the decreased precision resulting from using just one vignette rather than many, we might expect a smaller effect of aspect.

Another factor that can introduce variability among the observed effect sizes is the use of relatively small sample sizes. The original study had a relatively small sample, making the effect size estimate less precise than that of the RRR. All of the replication studies had larger sample sizes than the original study. Although it seems unlikely that sample size alone accounts for the observed differences, all of the effect sizes in the replication studies were numerically smaller than the ones reported in the original study.

Change in vignette

Another factor that might explain the difference in results is the change to one sentence in the vignette. With the agreement of the lead author of the original study (Hart), we changed "was pulling out his gun" to "pulled out his gun" in the imperfective aspect vignette. We were concerned that the original sentence might have engendered temporal disfluency because it implied that the character simultaneously drew and pointed his gun. This change meant that the imperfective aspect vignette included two rather than three instances of imperfective aspect and one instance of perfective aspect. The perfective aspect vignette included three instances of perfective aspect, as in the original study. Perhaps the reduction in the number of imperfective aspect actions reduced the power to observe an effect. Or perhaps the disfluency engendered by the original wording was necessary for the effect. However, the follow-up study conducted by Kurby and Kibbe used the original vignettes and found more intentionality attributed to the perfective aspect.

If the change in wording of the vignette explains the difference in results, then the original result either depends crucially on the number of imperfective actions or it resulted from a different mechanism (temporal disfluency) than originally thought. If either of these alternatives is true, then it is unclear whether the effect is robust enough to have an impact outside of the laboratory (e.g., in the justice system).

Conclusions

What can be concluded about the role of aspect in language processing and comprehension from this study? Linguistic analyses of aspect have typically focused on the role of aspect in conveying the temporal dynamics of events (Comrie, 1985; Madden & Ferretti, 2009; Vendler, 1957), and many studies show that aspect affects whether described events are perceived as ongoing or completed (Madden & Zwaan, 2003; Magliano & Schleich, 2000; Mozuraitis et al., 2013) and that aspect affects semantic activation of event knowledge associated with verbs (Ferretti et al., 2007; Ferretti, Rohde, Kehler, & Crutchley, 2009). These effects of aspect on the construction of mental models appear to be stable across studies. Understanding that an action was intentional follows from understanding that action in the first place. Therefore, one would expect effects of grammatical aspect on understanding action to be more stable than effects of aspect on understanding intentionality. The results of this RRR point in this direction. Exploring this idea more systematically would be fruitful for future research.

Appendix A: Vignette

After an argument broke out between James Westmoreland and Darryl McElroy in a 2009 dice game in East Cleveland, Westmoreland was pulling/pulled out his gun and was pointing/pointed it at Darryl McElroy. As the other players, including Darryl McElroy, attempted to run away, Westmoreland was firing/fired gun shots, one of which struck McElroy in the back, paralyzing him. McElroy and others identified Westmoreland as the shooter, and Westmoreland was later arrested and confessed to the crime.

Appendix B: Individual Lab Details

Amazon MTurk variant

Anita Eerland, Utrecht University
Andrew M. Sherrill, Northern Illinois University
Joseph P. Magliano, Northern Illinois University
Rolf A. Zwaan, Erasmus University Rotterdam
OSF: https://osf.io/z7kfe/

For the large-scale online experiment, participants were recruited from Amazon MTurk. They were paid $0.50 for participation and it took them about 10 min to complete the task. Participants signed up for a HIT that was either called "Forming impressions of others 1" or "Forming impressions of others 2." They were told they were about to read a story and answer questions about the characters. The approximate length of the experiment and the fact that the task required concentration were also mentioned. We only allowed MTurkers from the United States and those with a HIT approval rate >95% to participate.

We needed 130 participants for each version of the experiment. After a first round of collecting data, we ended up with 131 participants for the perfective aspect condition and 132 for

the imperfective aspect condition. Because we decided beforehand to exclude all non-native speakers of English, we excluded data from 1 participant in the perfective aspect condition and data from 7 participants in the imperfective aspect condition. Also, there were 4 participants who completed both versions of the experiment. For those participants, data of their second participation (based on the time log of their participation) were excluded. All these participants performed the perfective aspect version first. In total, we excluded data of 11 participants in the imperfective aspect condition. Then, we collected data from 9 additional participants in the imperfective aspect condition. Among these additional participants were 2 subjects that had already performed the task. Data of these participants were excluded and we collected data from 2 additional subjects. We ended up with 130 participants in both conditions. The final sample included 75 males (28.85%) and 185 females (71.15%). Age ranged from 18 to 76 (M = 39.00, SD = 12.69).

Lab studies

Jack D. Arnal, McDaniel College
OSF: https://osf.io/gdbrf/
Participants were recruited from the approved departmental participant pool. Those interested signed up for the experiment through the department's participant pool management software (Sona Systems) or via a Google Docs schedule. Instructions provided on both Sona and the Google Docs schedule invited participants to take part in a study about decision making. All other aspects of the study followed the prescribed protocol, including use of the provided Qualtrics script. Students who participated received partial course credit for their participation.

The goal was to have a minimum of 30 participants per condition, with initial analyses provided to the overall lead investigator of the replication project by November 15. Although the preregistered plan was to stop data collection on November 1, data collection was not terminated until November 14 (because sign-ups were slower than expected) with a total of 73 participants. The data from 6 participants were excluded from analyses because the participants reported as non-native speakers of English. The resulting sample sizes were 35 for the perfective aspect condition and 32 for the imperfective aspect condition. Of the 67 individuals included in the analyses, 17 reported as male (25.37%) and 50 reported as female (74.63%). The ages of participants ranged from 18 to 36 (M = 19.69, SD = 2.82).

Stephanie A. Berger, College of Mount Saint Vincent
OSF: https://osf.io/bcdfm/
Students completed the study in our psychology lab and earned extra credit in psychology courses for participating. A majority of our students are bilingual, so we excluded data from students whose second language was English and who estimated using English less than 90% of the time based on a short survey. Their data were eliminated from the file based on the IP address of the lab computer and the date and time they completed the study. Our goal was to have 40 native or primary English speakers in each condition. After running the first 108 participants, only 48 (44%) met the language requirement (n = 23 perfective, n = 25 imperfective). We continued recruiting in multiples of 8, collecting data from a total of N = 164 students. Of the 164 total participants, 82 were eliminated because of the language requirement—3 because of equipment problems and 4 because they completed the study twice (data from their first time in the study was included in the analysis). The final sample included n = 35 in the perfective aspect condition (9% male, age; M = 19.26, SD = 1.34) and n = 40 in the imperfective (15% male, age; M = 19.13, SD = 1.38). We did not meet our goal of 40 participants in each condition, but competition for our small subject pool prevented us from recruiting additional participants.

Angela R. Birt, Mount Saint Vincent University
Philip Aucoin, Mount Saint Vincent University
OSF: https://osf.io/gducj/
A total of 70 students from Mount Saint Vincent University in Halifax, Nova Scotia, participated in the study. They were recruited from undergraduate courses at the university, were tested in groups of 1–4 using Qualtrics software, and were paid in exchange for their participation. We excluded 4 participants from analyses: 3 were excluded because English was not their native language and 1 was excluded due to being younger than 18 years of age. Two participants were initially excluded due to what was originally considered as missing data, but they were reincluded as this was not the case. This resulted in a total sample size of 66 participants who met the inclusion criteria: for the perfective condition, n = 33, 75.80% female, age; M = 20.06, SD = 2.05, and for the imperfective condition: n = 33, 81.80% female, age; M = 21.30, SD = 5.06. Other than one of the primary student research assistants not participating in carrying out the RRR from the beginning, all procedures followed the approved protocol and did not deviate from our preregistered plan.

Anita Eerland, Utrecht University
Andrew M. Sherrill, Northern Illinois University
Joseph P. Magliano, Northern Illinois University
Rolf A. Zwaan, Erasmus University Rotterdam
OSF: https://osf.io/z7kfe/
Participants were 100- and 300-level undergraduate psychology students at Northern Illinois University (NIU). Participants were recruited through in-class announcements and an online bulletin board (Sona Systems). Each participant took approximately 10 min to complete all procedures and was compensated with course credit. The lab space included eight individual rooms, though no more than 5 participants completed the study at any given time. Each room had a Dell desktop on which study materials were administered via Qualtrics. Informed consent and debriefing were conducted with each participant individually.

Participants completed the study in the room alone and with the door closed. True random assignment was executed by flipping a coin before each participant entered the lab. The experimenter remained blind to study conditions by obfuscating the labels of study materials.

In total, 126 participants were recruited. Following preregistered exclusionary criteria, 18 participants were excluded for not being native English speakers and 8 participants were excluded for being over 25 years old. Data collection continued on an individual basis (within session blocks of up to 5 participants) until preregistered target sample sizes were achieved (50 per condition), accounting for exclusionary criteria. When 50 participants met inclusion criteria for the perfective condition, 46 participants currently met inclusion criteria for the imperfective condition. To balance conditions, 5 additional participants were recruited and assigned to the imperfective condition, with 1 excluded for being a non-native English speaker. In the final sample ($N = 100$; 50 per condition), 70 participants were female and 30 participants were male. The average age was 19.95 years ($SD = 1.46$).

Todd R. Ferretti, Wilfrid Laurier University
OSF: https://osf.io/5uf6s/
Participants were recruited from the Department of Psychology undergraduate testing pool by signing up online (Sona Systems). Participants were also recruited through posters placed around Wilfrid Laurier University. There were 86 participants in total. Thirty-eight of them were undergraduates that signed up through the departmental testing pool and received course credit for their participation. The first 11 undergraduate participants recruited through recruitment posters received $11 for their participation. However, due to the slow pace of recruitment, compensation was modified so that participants received $16 for participation in the study. As a result, a further 37 participants received $16 for their participation. One participant was removed for not meeting the criteria that participants had to be native English speakers. The data analysis was conducted on the remaining 85 participants, which included 43 in the imperfective condition and 42 in the perfective condition. The average age of the 55 female participants (65%) and 30 male participants (35%) was 20.09 years old ($SD = 2.81$).

The lab used consists of three separate rooms that contain a Mac desktop computer. Participants performed the study individually on the Mac computers in these rooms. The task took approximately 15 min to complete, including the time to read and sign the consent form.

Michael M. Knepp, University of Mount Union
OSF: https://osf.io/hxaq4/
Participants were undergraduate students recruited from psychology courses at the university. The Sona Systems research management system was used to recruit subjects and to ensure anonymity of the data. Within the Sona system, students were given a link to the study after sign-up and credit was automatically granted by the system following completion of the questionnaires. Student received .5 Sona credits for completing the online study. Random assignment to groups was done within the Qualtrics survey and each subject had an equal chance of being selected for either of the two conditions. Ninety students took part in our online-only version of the replication. Four students were excluded from the final analysis as they did not indicate English as their primary language. Within the 86 student sample, both groups had an equal gender ratio (13 men, 30 women) for a total of 26 men (30.2%) and 60 women (69.8%). There was no difference in age between the imperfective ($M = 19.16$, $SD = 1.09$) and perfective groups ($M = 19.33$, $SD = 1.41$, $p > .10$).

Christopher A. Kurby, Grand Valley State University
Mackenzie R. Kibbe, Grand Valley State University
OSF: https://osf.io/xiedk/
Participants were introductory-level undergraduate psychology students at Grand Valley State University. Participants were recruited through an online bulletin board (i.e., Sona Systems). Each participant took approximately 10 min to complete all procedures and was compensated with credit to satisfy course requirements. In total, 136 participants were recruited. Fourteen participants were excluded for not being native English speakers, and two participants were excluded because of missing data. In the final sample ($N = 120$), 89 participants were female and 31 participants were male. The average age was 18.62 years ($SD = 1.58$).

Data was collected by one undergraduate student. The psychology lab room had eight Dell desktops on which the surveys were administered via Qualtrics. A language history questionnaire was completed with paper and pencil. Debriefing was conducted with the participants as a group. Participants completed the study in the same room on tables with separators between them and with the door closed. At no point did participants interact with each other during the study. Participants were randomly assigned to condition using block randomization of 10-participant blocks to ensure an equal number of participants per cell. Sixty participants were assigned to each experimental condition (perfective and imperfective). The experimenter was blind to study conditions.

Joseph M. Melcher, St. Cloud State University
OSF: https://osf.io/3g8bh/
We recruited from the St. Cloud State University (St. Cloud, MN, USA) Department of Psychology participant pool, which consists of students taking a psychology course whose instructor offers extra credit for participating in studies. The pool is administered with Sona Systems, an online system through which students can browse and sign up for available studies. It also allows invitations to be sent on the basis of participant characteristic filters (e.g., self-reported native language is English). Our lab was aiming for 60 participants (30 per condition). Between September 8 and December 3, 2014, 69 students participated. All participants responded to all of the questions in the Qualtrics script. The data from 3 participants was excluded

because they self-reported a native language other than English as part of the demographic survey contained in the Qualtrics script. No other participants were excluded. This left 33 participants per condition. Participants received course extra credit based upon the allotted 30 min. Participants were run in groups of 1–3, each on a computer in separate rooms. The sample characteristics are consistent with our subject pool characteristics and the sample from the original Hart and Albarracín study: There were 18 males (27%) and 48 females (73%). Ages ranged from 18 to 50 (median = 20.0; M = 22.8; SD = 7.1).

Stephen W. Michael, Mercer University
OSF: https://osf.io/8y6bf/
Ninety undergraduates were recruited from introductory psychology courses at a private university in the Southeast in exchange for course credit. Sign-up sheets for a study on decision making were posted on a bulletin board in the psychology building. Consistent with preregistration exclusionary criteria, the only stated qualifications for participation were that the individual be 18 years of age and speak English as their primary language. The testing area was a computer lab in the psychology department where participants, in groups of 1–5, were unable to see others' computers screens. They also wore headphones throughout the study. Study materials (Qualtrics scripts) were administered on Dell desktop computers. Pseudorandom assignment was used whereby 45 participants were randomly assigned to Conditions 1 and 2. Research assistants were blind to the corresponding verb aspect conditions. Seven participants were excluded from analyses after indicating a language other than English as their primary language, leaving a final sample of 83 participants. Although preregistration plans were to recruit a minimum of 40 students in each condition, unequal exclusions across conditions resulted in 39 participants in the perfective condition and 44 in the imperfective condition. However, because preregistration for this lab specified a stopping point at 90 participants, no more participants were recruited. This final sample was 68.5% female with an average age of 18.71 (SD = 1.03).

Christopher Poirier, Stonehill College
Nicole Capezza, Stonehill College
Candace Crocker, Stonehill College
OSF: https://osf.io/px6n2/
We recruited participants from the psychology department's participant pool at Stonehill College. The participants were enrolled in one or more of the following courses: General Psychology, Developmental Psychology, Social Psychology, and/or Introduction to Statistics. They participated in the study as part of one option for course credit. We used a prescreening process in Sona Systems to recruit only participants who met the specified inclusion criteria (e.g., Is English your first and primary language?). A total of 80 participants completed the study; however, 1 participant was excluded because she did not follow instructions (i.e., the participant did not read the case study before advancing to the next part). The final sample consisted of 39 participants in the imperfective condition and 40 participants in the perfective condition. Candace Crocker served as the experimenter for every session, and she was blind to condition assignment. Our procedures followed the approved protocol and did not deviate from our preregistered plan.

Jason M. Prenoveau, Loyola University Maryland
Marianna Carlucci, Loyola University Maryland
OSF: https://osf.io/trxbd/
Participants were undergraduates recruited from the Psychology Research Pool. Participants received either course credit or extra credit for their participation in the study. Participants signed up to participate in the study at a given time using the Psychology Research Pool online recruitment tool.

When participants arrived at the laboratory, they were greeted by one of four research assistants. All four research assistants were trained in administering the study protocol and had run at least two pilot participants (whose data were not used for final analyses). All subjects completed the protocol in the same room. The room has five computers that are separated using dividers so that their screens are not visually accessible to individuals sitting adjacent to one another. Subjects were run both individually, and in small groups, depending on how many signed up for a given time slot; per the replication protocol requirements, these small groups did not exceed 5 participants. Subjects were instructed not to speak to one another during testing.

A random number generator (random.org) was used to assign participants to conditions; research assistants were blind to the conditions participants were assigned to. The target sample size was 50 participants per cell that meet the demographic criteria (i.e., native-English speakers between the ages of 18 and 25) and participants were run until this target was met.

In the perfective condition, 5 of the first 55 participants were excluded because 2 were outside of the specified age range and 3 did not identify English as their primary language. Because data were collected for 74 participants in the perfective condition, 19 additional participants were excluded because they exceeded the target sample size of 50 for this condition. In the imperfective condition, 1 of the first 51 participants was excluded because they were outside of the specified age range. Because data was collected for 53 participants in the imperfective condition, 2 additional participants were excluded because they exceeded the target sample size of 50 for this condition.

The perfective condition had an average age of 19.0 (SD = 1.2) and was 80% female. The imperfective condition had an average age of 19.0 (SD = 1.0) and was 74% female.

Issues arose during data collection that were not fully covered in the predata collection methods plan. First, there were 4 participants that had problems with Qualtrics and were not able to complete the study. Data from these participants were not recorded by Qualtrics and therefore these participants were not included in the analyses. Second, there was 1 participant who told the research assistant that they just clicked through and did not read the questions, and 2 others who research assistants

noted spoke during the experiment. However, because of the methods used in the present study (i.e., multiple participants run at the same time and no participant identification number given to participants), there was no way to determine which data corresponded to these participants. Thus, their data could not be excluded from possible inclusion in the analyses.

Acknowledgments

The authors would like to thank William Hart for his cooperation throughout the process and for his constructive feedback on the protocol for the replication studies. We thank Geoff Cumming for early feedback on the meta-analysis and Daniel Simons, Lidia Arends, and Samantha Bouwmeester for their input on performing the meta-(mediation) analyses. Also, we would like to thank Jelte Wicherts for feedback on a previous version of this article. Finally, we thank the following people (alphabetical order) for their help in conducting the experiments: Blake Alexi, Ashley Buck, Joseph Catalano, Molly Cioffi, Kaitlyn Fritz, Jessica Hagin, Jeffrey Hong, Yin Kong, Megan Lund, Kaitlyn Moriarty, Danielle Power, Carley Rampy, Caitlin Romano, and Miamoua Vang.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

Funding for participant payments was provided to individual labs by the Association for Psychological Science via a grant from the Center for Open Science.

Notes

1. These effect sizes that we calculated ourselves are slightly different from the effect sizes mentioned in the original paper.
2. The mention of "in-lab replication studies" includes the one online study run by Knepp.
3. τ is a measure of the total heterogeneity. I^2 is an estimate of the proportion of the heterogeneity that goes beyond what would be expected by chance. It is the total heterogeneity divided by the total variability. H^2 is the total variability divided by the sampling variability. The closer it is to 1, the more the variability across effect size estimates is consistent with sampling variability rather than meaningful heterogeneity. Q is a null-hypothesis test of whether there is meaningful heterogeneity.
4. T is an estimate of τ.

References

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology*, *51*, 1173–1182.

Carreiras, M., Carriedo, N., Alonso, M. A., & Fernandez, A. (1997). The role of verb tense and verb aspect in the foregrounding of information during reading. *Memory & Cognition*, *25*, 438–446.

Comrie, B. (1985). *Tense*. Cambridge, England: Cambridge University Press.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Ferretti, T. R., Kutas, M., & McRae, K. (2007). Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 182–196.

Ferretti, T. R., Rohde, H., Kehler, A., & Crutchley, M. (2009). Verb aspect, event structure, and coreferential processing. *Journal of Memory and Language*, *61*, 191–205.

Hart, W., & Albarracín, D. (2011). Learning about what others were doing: Verb aspect and attributions of mundane and criminal intent for past actions. *Psychological Science*, *22*, 261–266.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Communication Monographs*, *76*, 408–420.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

Kozak, M., Marsh, A. A., & Wegner, D. M. (2006). What do I think you're doing? Action identification and mind attribution. *Journal of Personality and Social Psychology*, *90*, 543–555.

Madden, C. J., & Ferretti, T. R. (2009). Verb aspect and the mental representation of situations. *The Expression of Time*, *3*, 217–231.

Madden, C. J., & Zwaan, R. A. (2003). How does verb aspect constrain event representations? *Memory & Cognition*, *31*, 663–672.

Magliano, J. P., & Schleich, M. C. (2000). Verb aspect and situation models. *Discourse Processes*, *29*, 83–112.

Morrow, D. G., Greenspan, S. L., & Bower, G. H. (1987). Accessibility and situation models in narrative comprehension. *Journal of Memory and Language*, *26*, 165–187.

Mozuraitis, M., Chambers, C. G., & Daneman, M. (2013). Younger and older adults' use of verb aspect and world knowledge in the online interpretation of discourse. *Discourse Processes*, *50*, 1–22.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. *Psychological Methods*, *7*, 422–445.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. A. (2012, October 14). *A 21 word solution*. (October 14, 2012). Retrieved from http://ssrn.com/abstract=2160588

Van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York, NY: Academic Press.

Vendler, Z. (1957). Verbs and times. *The Philosophical Review*, *66*, 143–160.

Young, L., & Waytz, A. (1993). Mind attribution is for morality. In S. Baron-Cohen, D. J. Cohen, & H. Tager-Flusberg (Eds.), *Understanding other minds: Perspectives from developmental social neuroscience* (pp. 93–103). Oxford, England: Oxford University Press.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language and memory. *Psychological Bulletin*, *123*, 162–185.