

CAUSAL ATTRIBUTION AND PRONOUN INTERPRETATION

Joshua K. Hartshorne

Harvard University

Running title: Causal attribution and pronoun interpretation

Key words: pronoun resolution, implicit causality, thematic roles, inference, pragmatics

Word count: 7554

Send Correspondence to:

Joshua Hartshorne

33 Kirkland Street

WJH 1120

Cambridge, MA 02138

Email: jharts@wjh.harvard.edu

Tel: 617-496-4486

Fax: 617-495-3728

Acknowledgements:

The author wishes to thank Jesse Snedeker, Manizeh Khan, Hugh Rabagliati and Rebecca Nappa for comments and discussion. Portions of this work were presented at the 2011 CUNY Conference on Human Sentence Processing and benefitted from discussion there. This material is based on work supported by a National Defense Science and Engineering Graduate Fellowship and the Gordon W. Allport Memorial Fund.

ABSTRACT

When asked who caused an event, causal attributions vary with the verb used (compare *Sally frightens Mary* with *Sally loves Mary*). Similarly, in causal dependant clauses, the preferred referent of a pronoun varies systematically with the verb in the main clause (contrast *Sally frightens Mary because she...* with *Sally loves Mary because she...*). Results from causal attribution studies are widely assumed to generalize to pronoun resolution studies and *vice versa*. Nonetheless, the primary evidence that the two phenomena are linked is that they share the same name: "implicit causality." This is potentially problematic as a significant literature has arisen around implicit causality, informing social psychology, developmental psychology and cognitive psychology -- particularly psycholinguistics. Four experiments demonstrate show little systematic relationship between the results of the two tasks, an outcome that generalizes across two different causal attribution and two different pronoun resolution tasks. Moreover, nonlinguistic world knowledge manipulations that affect causal attribution either do not affect pronoun resolution (Exps. 1 & 3) or have opposite effects on pronoun resolution (Exp. 2). These findings motivate a significant reanalysis of over three decades of findings across multiple subdisciplines within psychology.

Introduction

Who caused each of the following, Sally or Mary?

(1) Sally loves Mary.

(2) Sally frightens Mary.

On the surface, neither of these sentences provides any evidence one way or another. Nonetheless, numerous studies have shown that people credit Mary in first case and blame Sally in the second (see also Brown & Fish, 1983b; for a review, see Rudolph & Forsterling, 1997). Verbs like *love* are termed “object-biased,” in that people attribute causality to the verb’s object, whereas verbs like *frighten* are termed “subject-biased.”

Now, who does the pronoun *she* refer to in each of the following:

(3) Sally loves Mary because she...

(4) Sally frightens Mary because she...

Once again, there are no transparent clues, yet studies again show systematic behavior: people again choose Mary in the first case and Sally in the second (Garvey & Caramazza, 1974; Rudolph & Forsterling, 1997). Once again, there are numerous “object-biased” verbs like *love* – which induce people to resolve pronouns to the verb’s object – and numerous “subject-biased” verbs like *frighten* – which induce people to resolve pronouns to the verb’s subject.

Both phenomena are named “implicit causality,” and over the last three and a half decades, they have received a great deal of attention in a range of subdisciplines within psychology, particularly social psychology (Brown & Van Kleeck, 1989; Corrigan, 1988, 2001, 2002; Franco & Arcuri, 1990; Kasof & Lee, 1993; LaFrance,

Brownell, & Hahn, 1997; Maas, Salvi, Arcuri, & Semin, 1989; Mannetti & De Grada, 1991; Semin & Fiedler, 1988; Semin & Fiedler, 1991; Semin & Marsman, 1994) and psycholinguistics (Caramazza, Grober, Garvey, & Yates, 1977; Cozjin, Commandeur, Vonk, & Noordman, in press; Crinean & Garnham, 2006; Featherstone & Sturt, 2010; Ferstl, Garnham, & Manouilidou, in press; Fukumura & van Gompel, 2010; Garnham, Traxler, Oakhill, & Gernsbacher, 1996; Garvey & Caramazza, 1974; Garvey, Caramazza, & Yates, 1974; Greene & McKoon, 1995; Guerry, Gimenes, Caplan, & Rigalleau, 2006; Kehler, Kertz, Rohde, & Elman, 2008; Koornneef & Van Berkum, 2006; Long & De Ley, 2000; McDonald & MacWhinney, 1995; McKoon, Greene, & Ratcliff, 1993; Pickering & Majid, 2007; Pyykkonen & Jarvikivi, 2010; Stewart, Pickering, & Sanford, 2000). Both social and cognitive psychologists have studied the development of these phenomena in children (Au, 1986; Corrigan, 2003; Corrigan & Stevenson, 1994; Rudolph, 2008). In the context of the Sapir-Whorf hypothesis (Whorf, 1956), researchers have asked whether IC is an effect of language on thought or of thought on language (Brown & Fish, 1983a; Hoffman & Tchir, 1990).

Linking Causal Attribution and Pronoun Resolution: Evidence

Brown and Fish (1983b) named their causal attribution effect (1-2) "implicit causality" after the Garvey and Carmazza's (1974) pronoun resolution, itself an implicit hypothesis that the two phenomena were ultimately one and the same. This hypothesis has been widely adopted, with researchers rarely distinguishing between the phenomena in literature reviews and freely generalizing findings from one paradigm to "implicit causality" in general (Brown & Fish, 1983a, 1983b;

Corrigan, 1988, 2001, 2002; Goikoetxea, Pascual, & Acha, 2008; Greene & McKoon, 1995; Long & De Ley, 2000; Pickering & Majid, 2007; Rudolph, 2008; Rudolph & Forsterling, 1997). The clearest example of this implicit hypothesis is Rudolph and Forsterling's (1997) widely-cited landmark meta-analysis of implicit causality, which collapsed data across both paradigms.

This imprecision is of little import if the implicit hypothesis is correct and all "implicit causality" studies truly measure the same thing. Surprisingly, there is little or no evidence to support this hypothesis. No studies have directly compared the results of the two tasks, and the extant indirect evidence is quite weak.

Though this implicit hypothesis is typically asserted -- implicitly -- rather than defended, there are at least two types of indirect evidence one might marshal in favor of the implicit hypothesis. The first is that experiencer-subject verbs (verbs for which the subject experiences a mental state about the object: *Mary loves/hates/knows/understands Sally*) are typically object-biased in both types of tasks, while experiencer-object verbs (verbs for which the object experiences a mental state in relation to the subject: *Mary frightens/confuses/delights/fascinates Sally*) tend to be subject-biased in both types of tasks (e.g., Au, 1986; Brown & Fish, 1983b; Ferstl, et al., in press; Goikoetxea, et al., 2008). Even then, there are numerous exceptions to the overall pattern (Ferstl, et al., in press; Goikoetxea, et al., 2008), and there has been no systematic study of whether the exceptions are the same in both paradigms. Even in the best-case scenario, experiencer-subject and -object verbs account for fewer than three hundred out of thousands of English verbs (Levin, 1993). Whether other types of verbs show consistent patterns of bias is not

well-established, primarily because there is little consensus in the implicit causality literature as to how to identify other verb types, making comparison across studies difficult (Rudolph & Forsterling, 1997).

The second piece of indirect evidence is that changing semantic features of the subject and object of the verb can modulate the implicit causality bias for both task types (Corrigan, 1988, 2001, 2002; Ferstl, et al., in press; Garvey, et al., 1974; LaFrance, et al., 1997; Maas, et al., 1989; Mannetti & De Grada, 1991). That the same manipulations affect both phenomena could be taken as evidence that the same construct/mechanism underlies both phenomena. However, there is little evidence that the *same* manipulations affect both phenomena.

Several studies manipulated the gender of the subject and object. LaFrance et al. (1997) reported three causal attribution experiments, finding that (a) the subject is judged as more causal when the subject is male and the object is female (*John admired Mary*) than *vice versa* (*Mary admired John*), an effect which was stronger for negatively-valenced verbs, and similarly (b) the object is judged as more causal when the subject is male and the object female (*John admired Mary*) than *vice versa*, though this latter effect was less consistently observed across the three experiments.

Of three studies to investigate a similar manipulation in pronoun resolution tasks (Ferstl, et al., in press; Goikoetxea, et al., 2008; Mannetti & De Grada, 1991), only Ferstl et al. found an effect on pronoun resolution, reporting a slight overall male bias, with 52% of pronoun resolutions overall referring to male characters, an effect which was stronger for negatively-valenced verbs. LaFrance et al. (1997) did

not report overall verb biases, but calculations based on the table of results for Experiment 3 -- the experiment most directly applicable to Ferst et al.'s (in press) results -- suggest similarly that males are judged more causal for negatively-valenced verbs whereas females are judged more causal for positively-valenced verbs, though the effects are extremely small and unlikely to be significant (0.2 points on a 9-point Likert scale).

Several causal attribution studies have found effects of manipulating the social status (LaFrance, et al., 1997), potency (Corrigan, 2001, 2002), animacy (Corrigan, 1988), valence (Corrigan, 2002), and in-group/out-group status (Maas, et al., 1989) of the verbs' subjects and objects. There are no directly similar pronoun resolution studies, though Garvey et al. (1974) manipulated the "typicality" of the event participants for 26 verbs, reported results for 5 verbs, and found effects on only 3.

Thus, the only case in which the same manipulation has been reported to affect both causal attribution and pronoun resolution is gender, though in that case fully 2/3 of the pronoun resolution experiments found no effect, and the causal attribution results are complex and not easily compared to the one pronoun resolution experiment which did show an effect.

Linking Causal Attribution and Pronoun Resolution: Implications

Whether or not this linking hypothesis holds has considerable ramifications for theory. On a purely procedural level, given that discussions of both the causal attribution and pronoun resolution phenomena have drawn heavily on results from both phenomena, much of the literature will require careful reevaluation if the

hypothesis does not hold. For instance, much of the research into implicit causality has focused on whether implicit causality biases can be predicted based on the semantics of the verb (for review, see Rudolph & Forsterling, 1997). Much of this work has drawn largely or exclusively on causal attribution data (Au, 1986; Brown & Fish, 1983b; Rudolph & Forsterling, 1997; Semin & Fiedler, 1991).¹

There are broader theoretical implications as well. Researchers in the causal attribution literature have argued that implicit causality biases are the result of a complex inference about the typical causes of events, which takes into account gender roles, social hierarchies and other learned world knowledge. If indeed the implicit causality pronoun resolution biases are derived from these same underlying calculations, they are an illustration of linguistic processing taking as input an impressive array of nonlinguistic information. Recent studies demonstrate that people interpret pronouns consistent with the implicit causality bias by approximately half a second after pronoun onset (Cozjin, et al., in press; Koornneef & Van Berkum, 2006; Pyykkonen & Jarvikivi, 2010; Van Berkum, Koornneef, Otten, & Nieuwland, 2007; but see Stewart, et al., 2000), suggesting these complex inferences must be calculated extremely rapidly and are thus particularly strong evidence for the claim that complex pragmatic inferences emerge extremely rapidly

¹ Goikoetxea et al. (2008) and Ferstl et al. (in press) apply categorization schema derived from causal attribution studies to large sets of pronoun resolution data, but do not develop these schema further. In neither case do the schema provide a particularly close fit.

in language processing (e.g., Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Arnold, Hudson Kam, & Tanenhaus, 2007; Grodner, Klein, Carbary, & Tanenhaus, 2010; Kuperberg, Paczynski, & Ditman, in press; Sedivy, Tanenhaus, Chambers, & Carlson, 1999). In contrast, if the hypothesis linking the two implicit causality phenomena does not hold and it is found that world knowledge affects causal attribution but not pronoun resolution, that would place theoretically relevant limits on what information linguistic processing mechanisms are able to access.

In four experiments, I demonstrate that the hypothesis does not hold. Verbs may be subject-biased in causal attribution but object-biased in pronoun resolution. Manipulating the event participants' genders does not affect the two phenomena similarly, and manipulating their social status affects causal attribution but not pronoun resolution. In the General Discussion, I address the implications for theories of both implicit causalities, and of linguistic processing in general.

Experiment 1

Experiment 1 compares causal attributions and pronoun resolutions in order to test whether the former reliably predict the latter. Four types of verbs were used: causal verbs (*weakened, strengthened*), experiencer-object verbs (*bored, puzzled*), experiencer-subject verbs (*respected, hated*) and judgment verbs (*criticized, condemned*; see Appendix). According to Levin's (1993) linguistic analyses, the first two verb classes describe situations caused by their subject (*weaken = cause-to-be-weak; bore = cause-to-be-bored*) and the latter two describe situations elicited by their object (*respect = have-respect-elicited-by; criticize = have-criticism-evoked-by*). Thus, one may expect the first two to be subject-biased and the latter two to be

object-biased (Au, 1986; Greene & McKoon, 1995; Hartshorne & Snedeker, submitted).

Previous studies suggest that causal attribution is affected by the gender of the verb's subject and object, presumably due to effects of general world knowledge about gender roles (LaFrance, et al., 1997; Mannetti & De Grada, 1991). Thus, gender is manipulated in order to see whether pronoun resolution is similarly affected.

Method

Participants: Participants were tested via Amazon Mechanical Turk and compensated monetarily: 96 participants (95 native English speakers; 66 female; age 18-81, $M=36$, $SD=15$, one no report) in the causal-attribution task and 96 (92 native English speakers, one no report; 65 female, one no report; age 18-81, $M=37$, $SD=14$, two no report) in the pronoun-resolution task. 17 additional participants (13 in causal attribution) were excluded for failing to complete the task or for repeating the experiment.

Materials: Six verbs from each of the four classes of verbs were chosen. Subjects and objects of the verbs were chosen from common male and female names. Four stimulus lists were constructed, crossing trial order and whether the subject of the verb was male or female. On each list, half the subjects were male and half female.

Procedure: The causal-attribution task (5) was adapted from Experiment 1 in Brown and Fish (1983b).²

(5) Christopher affected Ashley. How likely is this because:

- a. Christopher is the kind of person who affects people.

not likely 1 2 3 4 5 6 7 8 9 definitely likely

- b. Ashley is the kind of person whom people affect.

not likely 1 2 3 4 5 6 7 8 9 definitely likely

The procedure for the pronoun-resolution task (6) was a modification of a sentence completion task often used to assess pronoun biases (e.g., Kehler, et al., 2008):

(6) Which word is the most likely continuation for the following sentence?

Christopher affected Ashley because

- a. he b. she

Results

Unless otherwise specified, analyses employed mixed-effects models with subjects and items as random effects in R using the lme4 package, and p-values were estimated using the function `pvals.fnc`, which implements Markov chain Monte Carlo sampling (see Baayen, 2008; Bates & Sarkar, 2007; R-development-core-team, 2005).

² Brown & Fish included a third question in each trial: *c) some other reason*.

However, responses to this question are rarely analyzed in the literature. In order to shorten the experiment, it was dropped.

Causal Attribution: There was no interaction between subject and object causal attributions and whether the subject was male and object female or vice versa, either overall ($t=1.21, p=.23$) or for any of the four verb classes analyzed individually ($ps>.23$). Subsequent analyses collapsed across this factor. All four classes of verbs were subject-biased ($ps<.001$; Figure 1), except experiencer-subject verbs, which showed no bias ($t=1.56, p=.12$).

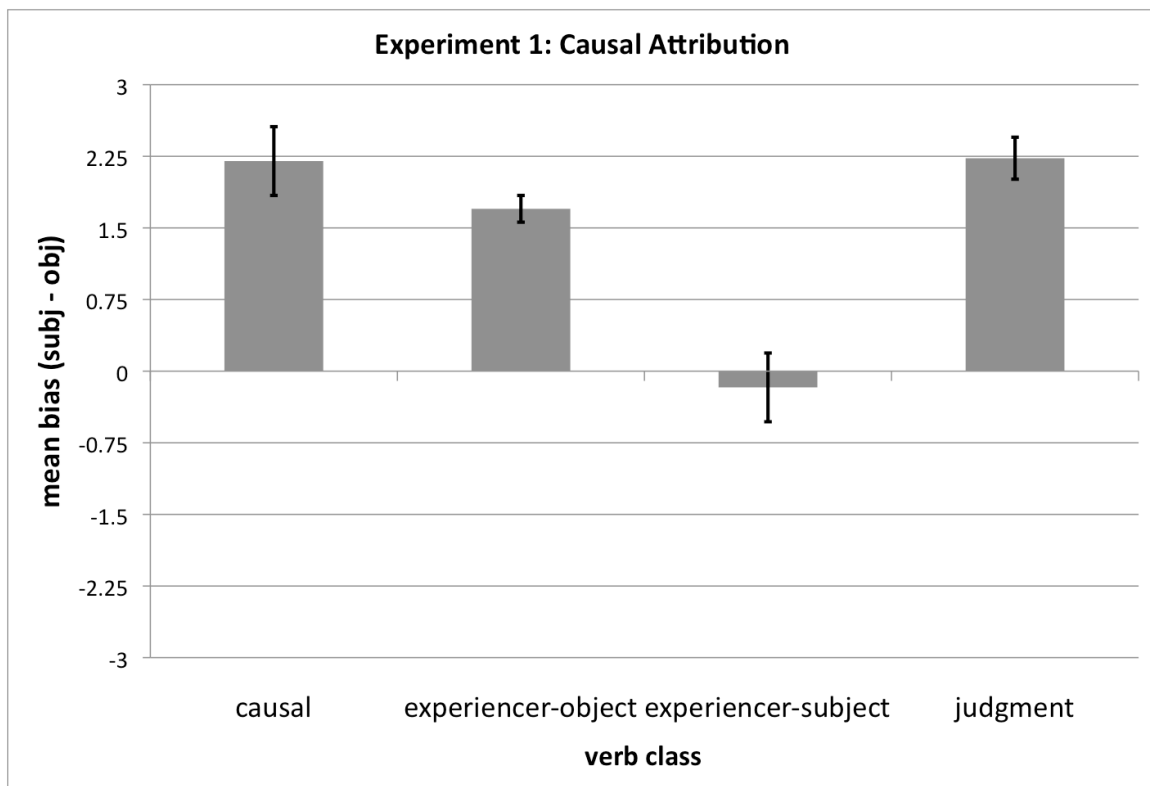


Figure 1. Mean causal-attribution biases (attribution to subject – attribution to object) for the four verb types, with standard errors, in Experiment 1. Here and throughout, standard errors are calculated by item.

Pronoun Effect: Again, gender had no effect, either overall ($t=.55, p=.58$) or in any of the four verb classes individually ($ts < 1, ps > .25$). Subsequent analyses collapsed across this factor. Participants preferred subject-referring pronouns to object-referring pronouns for causal verbs and experiencer-object verbs (respectively: $t=3.16, p=.001$; $t=11.52, p<.001$; Figure 2). In contrast, experiencer-subject verbs and judgment verbs were object-biased (respectively: $t=21.24, p<.001$; $t=16.45, p<.001$).

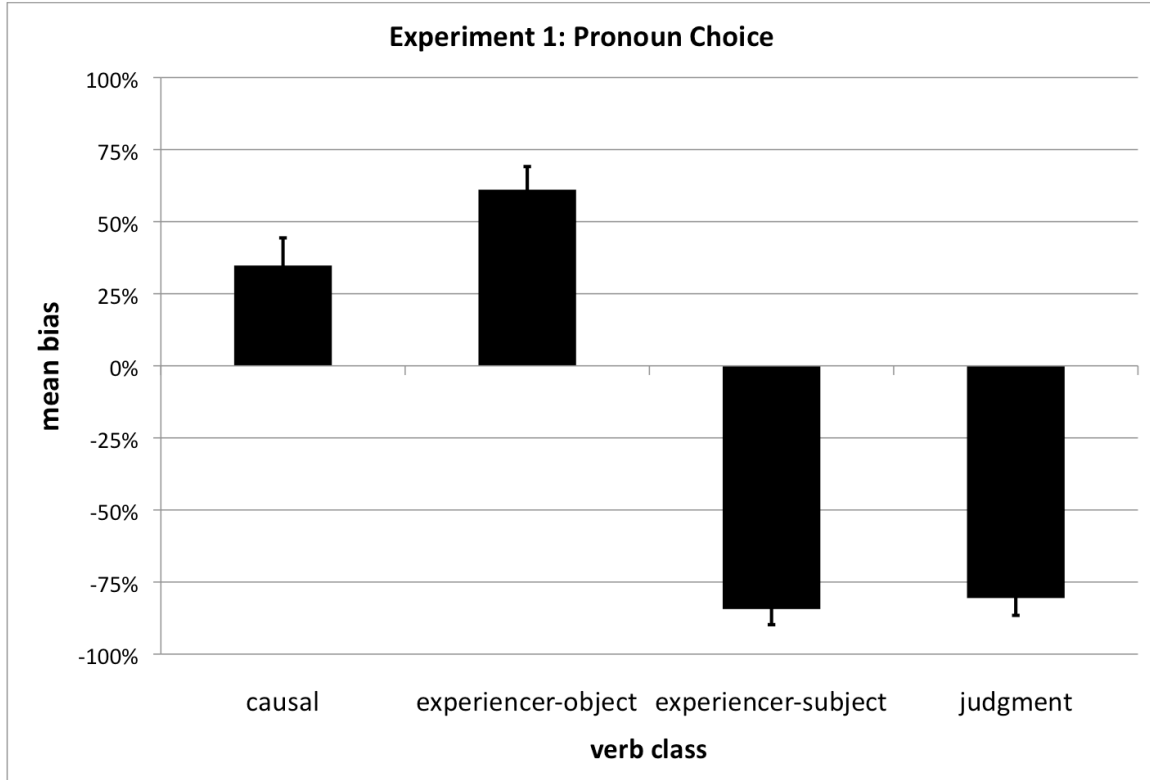


Figure 2. Average subject bias for the pronoun task in Experiment one by verb class, with standard errors. 100% = all participants chose subject-referring pronoun; -100% = all participants chose object-referring pronoun.

Discussion

While experiencer-object and causal verbs were subject-biased in both tasks, experiencer-subject verbs (*respect, hate*) were object-biased in pronoun resolution but unbiased in causal attribution. Judgment verbs (*criticize, condemn*) were numerically *more* subject-biased than experiencer-object and causal verbs in causal attribution but as object-biased as experiencer-subject verbs in pronoun resolution (no significant difference: $t < 1$). These results casts doubt on the claim that pronoun resolution relies on causal attribution (Brown & Fish, 1983b; Rudolph & Forsterling, 1997) and that results from one paradigm generalize to the other. Why causal attribution and pronoun resolution do not pattern together will be returned to in the General Discussion.

Experiment 2

In contrast with Ferstl et al. (in press), but in line with Mannetti and De Grada (1991) and Goikoetxea et al. (2008), Experiment 1 did not find that participants were more likely to expect pronouns to refer to the male character. With three studies showing no effect and only one which shows a very small effect, one may be tempted to discount Ferstl et al.'s results as a false positive. However, it is perhaps relevant that Ferstl et al. (in press) did not find the effect for all verbs, but rather the effect was particularly true of negatively-valenced verbs. Similarly, the one causal attribution study to investigate the effect of gender likewise found gender effects restricted to negatively-valenced verbs (LaFrance et al., 1997). Perhaps Experiment 1 simply included too many positively-valenced verbs.

Experiment 2 replicated Experiment 1, but with a subset of verbs taken from Ferstl et al. (in press). Thus, Experiment 2 tests whether Ferstl et al.'s findings with regard to gender replicate using the same verbs, and also tests whether they generalize to causal attribution. Ferstl et al. (in press) tested 96 participants. In order to increase the chances of replicating their (small) gender effect, a larger number of participants were included in Experiment 2.

Participants: Participants were tested via Amazon Mechanical Turk and compensation monetarily: 120 participants in the causal attribution task (116 native English speakers; 76 female; age 18-82, $M=37$, $SD=14$) and 120 participants in the pronoun-resolution task (110 native English speakers, 3 no response; 81 female; age 18-68, $M=37$, $SD=12$). An additional 69 participants (45 in causal attribution) were excluded -- most for failing to answer all items, and some for repeating the experiment or for low accuracy on filler items (see below).

Materials: 20 verbs were chosen from Ferstl et al. (in press): the 10 most negatively-valenced transitive verbs (according to their ratings) which also showed a numeric male bias (i.e., more attributions to the male character) and their 10 most positively-valenced transitive verbs which also showed a numeric female bias (i.e., more attributions to the female character). Four lists were made, counterbalanced as in Experiment 1 (i.e., with male-/female-biased as the verb classes).

In addition, four filler trials were created, with the first two and last two trials on each list being filler. Both characters on the filler trials were of the same gender (2 male, 2 female trials), rendering the pronoun resolution task unambiguous (*Christina believed Melissa because... she/he?*). For the causal

attribution task, these sentences were adjusted as to render causality unambiguous (*Christina believed Melissa because she was very gullible*), with the correct answer being the subject twice and the object twice. Participants who did not answer all fillers correctly were replaced.³

Procedure: The procedure was identical to that of Experiment 1.

Results

Causal Attribution: Subject-bias was calculated as in Experiment 1. Male bias for each verb was calculated in terms of how much more strongly the verb was subject-biased when the subject was male than when the subject was female. Overall, there was no strong male bias ($M=0.0$, $SE=0.1$; $t<1$). However, there was a tiny but significant ($t=3.63$, $p<.001$) trend for stronger male biases among the positively-valenced verbs ($M=0.3$, $SE=0.1$) than for the negatively-valenced verbs ($M=-0.3$, $SE=0.2$). It should be noted that these effects are small: 0.6 points on a 9-point Likert scale.

Pronoun Resolution: Subject-bias was calculated as in Experiment 1. Male bias for each verb was calculated in terms of how much more strongly the verb was subject-biased when the subject was male. Only half (5) of the negatively-valenced

³ Calculating the "correct" answer for causal attribution fillers was not trivial. The following algorithm was employed: Participants must have on average attributed more causality to sentence subject when that was the correct answer than when it was the incorrect answer and never have attributed more causality to the sentence subject when it was the incorrect answer than when it was the correct answer.

verbs -- which were numerically male-biased in Ferstl et al. (in press) -- were found to be numerically male-biased. Similarly, only half of the positively-valenced verbs found by Ferstl et al. (in press) to be female-biased were female-biased in the present results. Nonetheless, there was a barely significant ($t=2.01$, $p=.04$) stronger male bias for the negatively-valenced verbs ($M=21.0\%$, $SE=28.0\%$; SEs here and throughout calculated by verb) relative to the positively-valenced verbs ($M=6.6\%$, $SE=35.2\%$). Overall, there was a weak male bias ($M=13.8\%$, $SE=22.0\%$; $t=3.87$, $p<.001$).

Experiment 2: Causal Attribution vs. Pronoun Resolution

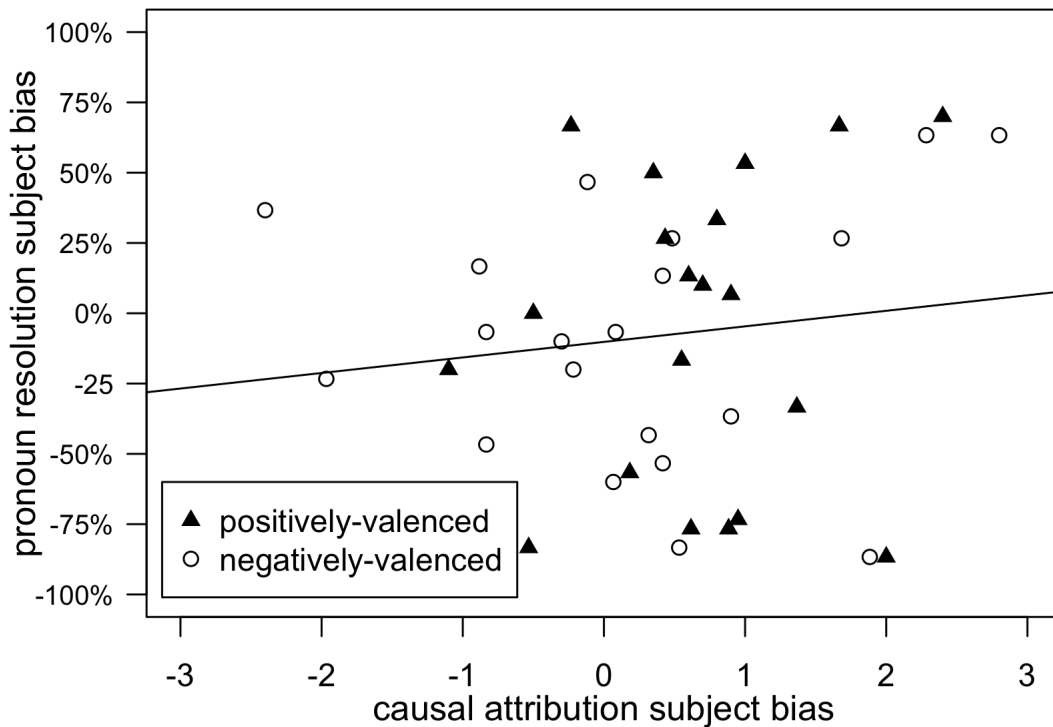


Figure 3. Correlation between causal attribution and pronoun resolution biases in Experiment 2, with male-subject and female-subject stimuli treated separately. Collapsing across the gender manipulation does not affect the pattern of results.

Combined: There was no correlation between subject-bias in the two tasks, whether the male-subject/female-object and female-subject/male-object versions of each sentence were treated as separate items ($r=0.12, p=.45$; Figure 3) or the same item ($r=-0.07, p=.77$). However, as noted above, negatively-valenced verbs were more male-biased than positively-valenced verbs in pronoun resolution, whereas the reverse was true in causal attribution. This observation is further bolstered by a significant negative correlation between male-bias on the two tasks ($r=-0.50, p(18)=.02$).

Discussion

The results of Experiment 2 do not support the claim that causal attribution and pronoun resolution implicit causality tasks measure the same thing. There was no correlation between biases elicited by the two tasks, and to the extent that the gender manipulation affected either, it did so in opposite directions.

Although the claim that causal attribution and pronoun resolution are equivalent tasks seems increasingly untenable, there are reasons to not abandon it too quickly. While there is little if any evidence for the proposal, it is nonetheless deeply embedded in the implicit causality literature, and considerable reanalysis of the literature will be necessary if the proposal is indeed false. Thus, I replicate the

non-equivalence of the two phenomena, using a different pronoun task (Experiment 3) and a different causal attribution task (Experiment 4).

One main theoretical consequence of the proposal was to support a view on which the pronoun resolution bias is an inference derived from world knowledge, the evidence being that (a) the causal attribution task appears to be a world knowledge task, and (b) the causal attribution task is known to be affected by world knowledge. Even if nothing can be generalized from causal attribution to pronoun resolution, that does not mean that the pronoun resolution bias is *not* a world knowledge inference. In fact, in Experiment 2, I found a small effect of gender on biases elicited by some negatively-valenced verbs, replicating Ferstl et al. (in press).

However, one must be cautious in interpreting this result. While regression to the mean would predict that retesting should lead to a weakening of the effect, fully half the verbs actually reversed directions in terms of their gender bias. This is what one expects only if the true mean is very close to or equivalent to 0% male bias. Consider, moreover, that each stimulus (*Sally plagued Mary because...* vs. *Mary plagued Sally because...*) was observed by 48 participants in Ferst et al. (in press) and 60 participants in Experiment 2, which should have provided quite accurate estimates of the male-bias for each verb. Instead, it appears that there is little test-retest reliability. In short, the effect was small, barely replicable, and apparently restricted to a small number of verbs. Thus, while these data show that in some cases pronoun resolution can be affected by world knowledge such as typical gender roles, they are not the sort of result around which one builds a theory.

Nonetheless, it may be that the gender manipulation was relatively weak, particularly among contemporary American speakers of English. Thus, in Exps. 3-4, I employ a more robust manipulation of non-linguistic world knowledge.

Experiment 3

Several studies have shown that causal attribution is affected by the social roles of the verbs' subjects and objects (Corrigan, 2001, 2002; LaFrance, et al., 1997). For instance, *The mugger struck the passerby* is subject-biased while *The passerby struck the mugger* is object-biased, presumably because muggers are more likely to both initiate and deserve physical blows (Corrigan, 2001). Perhaps this stronger manipulation will affect pronoun resolution.

Method

Participants: Participants were tested via Amazon Mechanical Turk and compensated monetarily: 48 participants (46 native English speakers, 1 no response; 33 female, one no response; age 18-67, M=38, SD=13, one no response) in the causal-attribution task and 48 participants in the pronoun-resolution task (46 native English speakers; 32 female; age 18-81, M=39, SD=15). An additional 23 participants (15 in causal attribution) were excluded for failing to complete all items or for repeating the task. Two additional participants that were not needed for counterbalancing were not analyzed.

Materials: The same verbs and lists from Experiment 1 were used. Six pairs of characters with defined social hierarchies (e.g., *duke* and *butler*, *king* and *knight*) were chosen. Each was assigned randomly to one verb from each of the four classes

(these roles were repeated across verb classes in order to aid direct comparison of verb class). Which character was the subject was counterbalanced across the lists.

Procedure: The procedure for the causal-attribution task was identical to Experiment 1. The procedure for the pronoun task is shown in (7):

(7) The butler blamed the duke because he is a froom.

Who is a froom? the butler the duke

Each sentence ended with a unique novel word such as *froom*. This procedure, introduced by Hartshorne and Snedeker (submitted), mitigates the fact that the material following the pronoun can override the pronoun bias and force particular resolutions (e.g., *Sally frightened Mary because she was easily scared*).

Results

Causal Attribution: Across all verbs, there was an interaction between subject and object causal attributions and social hierarchy, with sentences being more subject-biased when the subject was high-ranking ($t=7.22, p<.001$). Analyzing verb classes separately, this effect (Figure 4) was significant for judgment verbs ($t=6.06, p<.001$), experiencer-object verbs ($t=2.69, p=.008$) and experiencer-subject verbs ($t=5.78, p<.001$), but not for causal verbs ($t<1, p=.71$).

Collapsing across the hierarchy manipulation, causal verbs, experiencer-object verbs and judgment verbs were all significantly subject-biased (respectively: $t=8.90, p<.001$; $t=5.13, p=.006$; $t=2.02, p=.04$). Unlike in Experiment 1, experiencer-subject verbs were significantly object-biased ($t=8.61, p<.001$), an effect primarily visible with high-ranking objects (Figure 4).

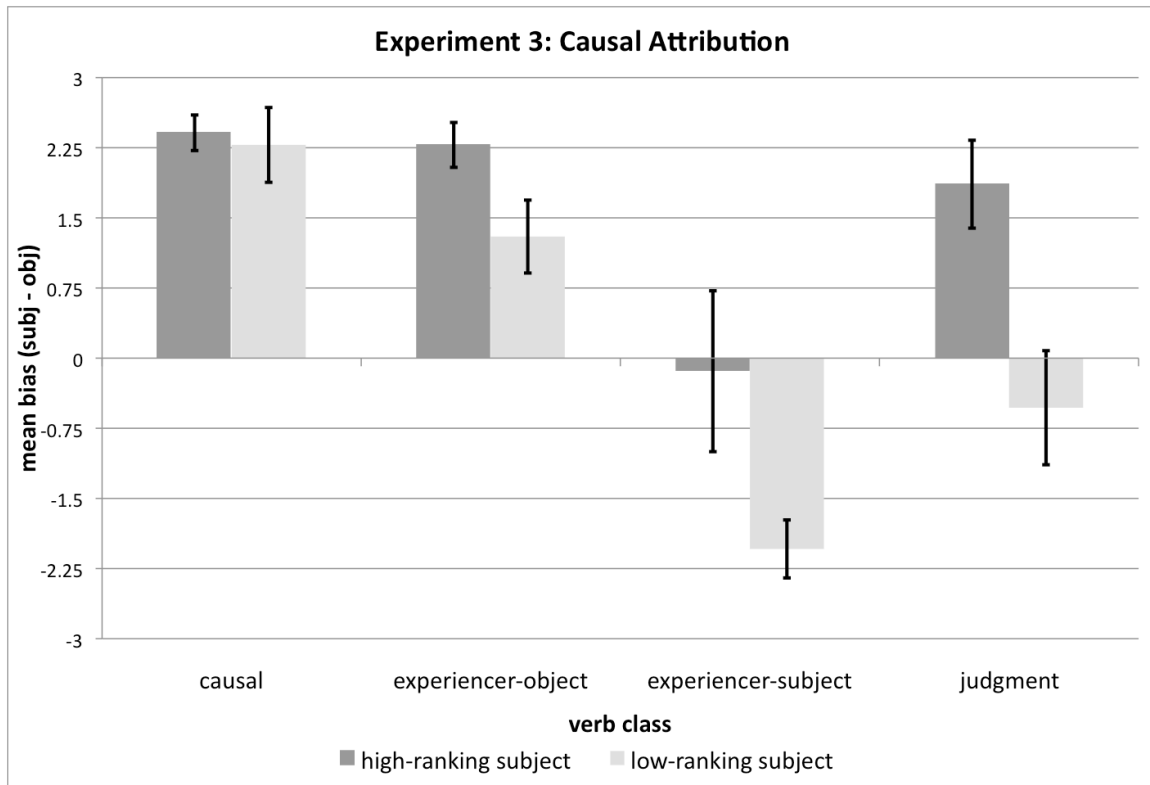


Figure 4. Mean causal-attribution biases (attribution to subject – attribution to object) for the four verb types, with standard errors, in Experiment 3.

Pronoun resolution: Pronoun resolution (Figure 5) was unaffected by whether the subject was high-ranking or low-ranking, either across all verbs or within each verb class analyzed individually ($t_s < 1$). Collapsing across the hierarchy manipulation, causal verbs and experiencer-object verbs were significantly subject-biased (respectively: $t=13.33$, $p < .001$; $t=17.22$, $p < .001$), while experiencer-subject verbs and judgment verbs were significantly object-biased (respectively: $t=3.80$, $p < .001$; $t=5.40$, $p < .001$), as in Experiment 1.

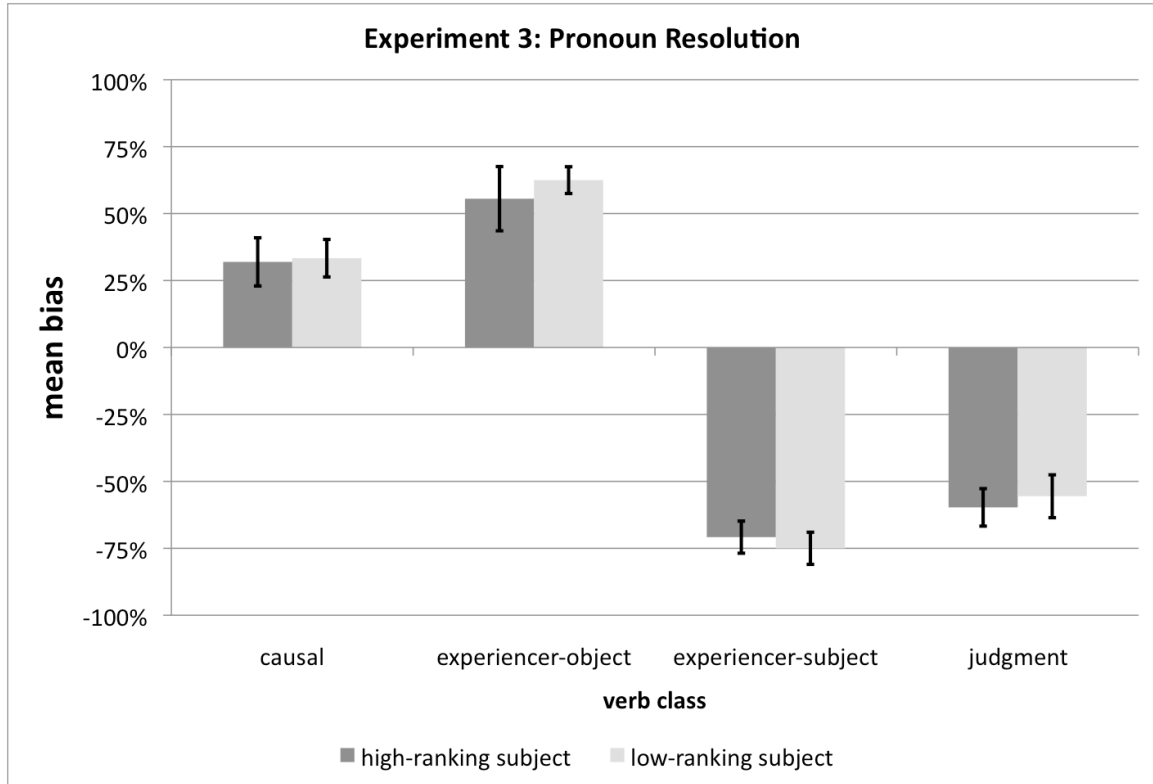


Figure 5. Mean subject biases for the pronoun-resolution task in Experiment 3, with standard errors. 100% = all participants chose subject-referring pronoun; -100% = all participants chose object-referring pronoun.

Discussion

As in Experiment 1, judgment verbs were subject-biased according to causal attribution but object-biased according to pronoun resolution. Thus, there appear to be systematic differences between the results of the two tasks. Moreover, the social hierarchy manipulation produced large effects on causal attribution but had no effect on pronoun resolution.

Experiment 4

Multiple measures have been used to assess causal-attribution bias, and there has been no careful investigation of whether the different methods provide the same results. Although the method used in Exps. 1-3 is the oldest and most widely-used measure, it is a very indirect probe of causality. Perhaps another measure would show a clear relationship between causal attribution and pronoun resolution. Experiment 4 replicates the causal-attribution task from Experiment 3, using a different paradigm.

Method

Participants: 48 participants (45 native English speakers; 29 female, 2 no response; age 18-62, $M=37$, $SE=14$, 2 no response) completed Experiment 3 via Amazon Mechanical Turk and received token financial compensation. An additional 10 participants were excluded for failing to complete all items or repeating the experiment.

Materials and Procedure: Materials and procedure were identical to the causal-attribution task used in Experiment 3 with the change that participants were asked to choose the more causally-responsible character:

(8) The butler blamed the duke.

Who is most likely responsible for this?: the butler the duke

Results and Discussion

Results are shown in Figure 6. Only experiencer-subject verbs showed a significantly stronger subject-bias when the subject was high-ranking ($t=4.82$, $p<.001$; all other $ps>.05$). Overall, causal verbs and experiencer-object verbs were significantly subject-biased (respectively: $t=25.54$, $p<.001$; $t=9.30$, $p<.001$) and

experiencer-subject verbs were significantly object biased ($t=2.44, p=.01$), at least when the subject was low-ranking (Figure 6). Crucially, judgment verbs showed no significant bias ($t=.11, p=.91$), though they were strongly object-biased in terms of pronoun resolution in Exps. 1 & 3.

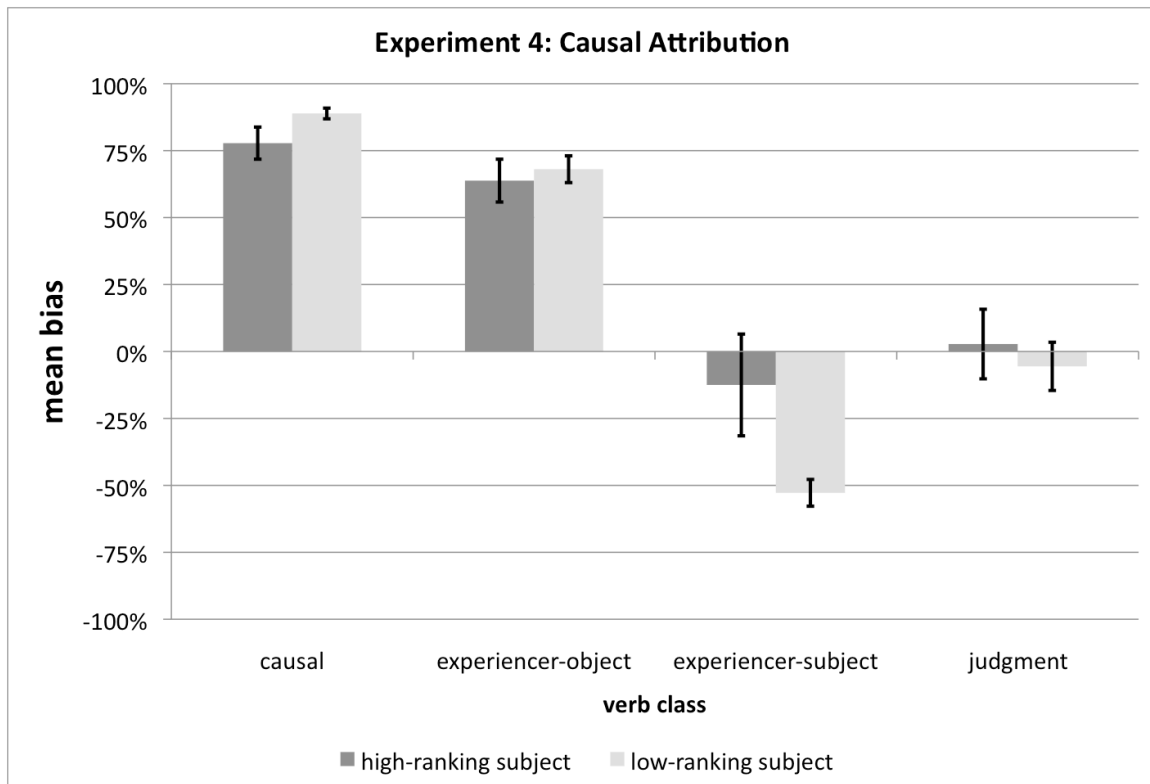


Figure 6. Mean causal attributions in Experiment 4, with standard errors. 100% = all participants chose subject as most likely responsible; -100% = all participants chose object as most likely responsible.

Qualitatively, there were two main differences between the results of the causal attribution tasks used in Experiments 3 and 4. First, the effect of the social hierarchy manipulation disappeared for all but experiencer-subject verbs, and in

fact reversed numerically for causal and experiencer-object verbs. In addition, judgment verbs were no longer biased, an effect driven primarily by the disappearance of the subject-bias in the high-ranking subject condition. This is particularly interesting in that judgment verbs were strongly subject-biased in Experiment 1, which did not have a social hierarchy manipulation. This suggests that not only are there systematic differences between the results of causal attribution and pronoun resolution tasks, different causal attribution tasks produce different results. The task used in Experiment 4 more clearly involved actual causal judgments. However, the task used in Exps. 1-3 is the most common in the literature, suggesting that much of the literature on causal attribution may not be measuring causal attribution as faithfully as has been supposed. These issues are returned to in the general discussion.

General Discussion

By calling their causal-attribution effect “implicit causality,” the same name given by Garvey and Caramazza (1974) to an apparently similar pronoun-resolution effect, Brown and Fish (1983b) were making an implicit claim that the two phenomena were the same. This claim has been taken at face value in the literature since (e.g., Rudolph & Forsterling, 1997) but never tested.

In a series of four experiments, I was unable to find any clear support for this claim. Rather, there were systematic differences in the data elicited from the two types of tasks. Many verbs that were object-biased according to pronoun resolution were unbiased or even subject-biased according to causal attribution. The social hierarchy manipulation employed in Experiment 3 affected causal attribution but

not pronoun resolution. To the extent that gender interacts with implicit causality biases, it does so in opposite directions for the two tasks (Experiment 2).

Much of the implicit causality literature is predicated on the assumption that the causal attribution and pronoun resolution tasks measured the same thing, and as such theoretical reviews have generally failed to distinguish between results stemming from one paradigm or the other (e.g., Rudolph & Forsterling, 1997). As the two tasks are in fact not equivalent, a careful reanalysis and re-synthesis of the literature, systematically distinguishing which results are known to be true of which task type, is required.

For instance, as discussed in the introduction, nearly all evidence that implicit causality is affected by world knowledge stems from causal attribution tasks (Corrigan, 1988, 2001, 2002; LaFrance, et al., 1997; Maas, et al., 1989; Semin & Fiedler, 1988; Semin & Fiedler, 1991; Semin & Marsman, 1994). There is very little evidence that pronoun resolution biases are similarly affected. Garvey et al. (1974) reported an effect restricted to three verbs. Similarly, Ferstl et al.'s (in press) gender effect appears to be quite small and specific to certain verbs (see also Experiments 1-2, above). Experiment 3 showed no effect of social hierarchy. Thus, there may be some small effects for some verbs, but world knowledge does not appear to be playing a defining role in pronoun resolution implicit causality tasks.⁴

⁴ Even if a world knowledge manipulation were to be found that robustly affected ultimate pronoun resolution, that does not necessarily imply that the pronoun resolution biases under investigation here are themselves derived from world

Similarly, work which has attempted to find semantic verb classes that predict implicit causality bias is based primarily on causal attribution (Au, 1986; Rudolph & Forsterling, 1997; Semin & Fiedler, 1988; Semin & Fiedler, 1991; Semin & Marsman, 1994; but see McKoon et al., 1993). Two large studies of pronoun resolution biases (Ferstl, et al., in press; Goikoetxea, et al., 2008) classified verbs according to Rudolph and Forsterling's (1997) criteria, finding many exceptions to the predicted patterns. This may be at least in part due to the fact attested in the present study: that the same verbs may be systematically classified differently according to causal attribution and pronoun resolution. Thus, future work into verb classifications will need to consider causal attribution and pronoun resolution separately.

knowledge. In *John frightened Mary because she...*, most people resolve *she* to Mary, even though *frightened* is strongly subject-biased. Nonetheless, overriding this bias comes at a processing cost (Caramazza, et al., 1977; Koornneef & Van Berkum, 2006; Stewart, et al., 2000; Van Berkum, et al., 2007). This suggests that the pronoun resolution bias exists independently of other factors that may contribute to determining ultimate pronoun resolution. One could study whether the pronoun resolution bias itself depended on world knowledge factors -- rather than merely being overridden by them -- if one could identify enough implicit causality sentences where such factors had a reliable effect. The fact that such sentences are difficult to find is data in and of itself.

Perhaps the most serious challenge -- to which the rest of this discussion is devoted -- is explaining *why* causal attribution and pronoun resolution diverge. This is well beyond the scope of the present study, as answering the question will likely require a number of studies itself. Here I merely sketch some possibilities that may serve as a guide to further research.

Causal Attribution vs. Pronoun Resolution

The most obvious reason for the divergence between causal attribution and pronoun resolution is the often overlooked fact that *neither* task directly probes causality or causal reasoning. Although both tasks crucially involve the word *because*, the word *because* introduces an explanation, not a cause (Kehler, 2002; McKoon, et al., 1993). Naturally, explanations typically do refer to causes, but already a free parameter has been introduced.

Moreover, events often have more than one (type of) cause, and people's explanations of events focus on different types of causes under different circumstances (for a review, see Lombrozo, 2010). *Mary criticized Sally* has at least two causes: Mary's voluntary initiation of the criticism event, and Sally's prior criticism-evoking action. Thus, one may choose to explain the criticism in terms of Mary's voluntary initiation *or* Sally's prior action (or in terms of something else). Different implicit causality tasks may well elicit different types of explanations. Both the subjects and objects of judgment verbs and subject-experiencer verbs seem to bear causal responsibility, whereas this seems less true of the causal verbs and experiencer-object verbs. This could explain why the former display more variable behavior across tasks than do the latter.

With these facts on the table, it should be noted that Brown and Fish's task adds some additional constraints not present in pronoun resolution tasks. Brown and Fish's task asks whether, for instance, *Mary criticized Sally* can be explained in terms of very specific properties of Mary and Sally: Mary being the type of person who criticizes people or Sally being the type of person who people criticize. To put this in perspective, if Mary broke a vase, it would be incontrovertible that she caused the breakage (barring demonic possession, etc.), though this might not be because Mary is the kind of person who breaks vases. She may be an extraordinarily careful person, and this an extraordinarily fragile vase.

These considerations may help explain why the Brown and Fish task is particularly susceptible to world knowledge manipulations. Decisions as to whether *the duke criticized the butler* is due to the duke being the kind of person who criticizes people may be contaminated by people's baseline expectations about how likely it is that dukes are critical people, quite independent of the fact that the duke criticized the butler. Interestingly, Experiment 4, which simply required that participants decide which character was responsible for the event, showed a much-reduced effect of world knowledge. In fact, the studies which have shown effects of world knowledge on causal attribution have employed Brown and Fish's task or a variant that similarly queried specific properties of the subject and object, rather than simply asking who was responsible/causal (Corrigan, 1988, 2001, 2002; LaFrance, et al., 1997; Maas, et al., 1989; but see Corrigan, 1988, Exp. 3).

Encoded Knowledge

If implicit causality biases -- particularly with respect to pronoun resolution -
- are not based on inferences from world knowledge, what *are* they based on?

Although the world knowledge hypothesis itself in many ways derives from Brown and Fish (1983b), so does an alternative: the biases are derived from the literal, encoded meaning of the verb. The world knowledge hypothesis likewise derives implicit causality biases from meaning, but from meaning that has been inferred rather than from the encoded meaning. The distinction between encoded and inferred meaning can be cashed out as one of entailment: words entail certain things to be true and may only imply others. If Sally frightened Mary, she may have done so by wearing a scary costume, shouting boo, telling a scary story, or in any number of other ways. While there can be variability in manner, one thing is invariant: Mary must be frightened. If she is not, then Sally did not frighten Mary.

In a recent proposal, Hartshorne and Snedeker (submitted), following analysis in the linguistic semantics literature (Jackendoff, 1990; Levin, 1993; Pesetsky, 1995), suggest that the causal information relevant to implicit causality actually forms part of the encoded meaning of the verb. For instance, *Sally frightened Mary* literally means *Sally caused Mary to feel fright*. Thus, the parser need not take into account any inferences about causality when determining pronoun resolution: the cause is actually explicitly encoded in the verb.

This account builds on a long tradition in the literature. Starting with Brown and Fish (1983b), a number of researchers have attempted to classify verbs according to their meaning in order to predict implicit causality biases (Au, 1986; Rudolph & Forsterling, 1997; Semin & Fiedler, 1991). Brown and Fish suggested

that action verbs were subject-biased, whereas experiencer-subject state verbs were object-biased and experiencer-object verbs were subject-biased.⁵ Brown and Fish's proposal depended on encoded meaning in order to classify verbs. However, they appear to view causality as inferred from this encoded meaning, rather than actually being encoded. In contrast, McKoon et al. (1993) argue that in all subject-biased verbs, the subject fills the semantic role of the *initiator* and the object, the *reactor*, whereas the reverse is true for all object-biased verbs. They argue that "a *because* clause will naturally then explain what property or action of the initiator provoked the response by the reactor" (p. 1042). Thus, on their account, implicit causality is closely tied to the literal meaning of the verb.

If implicit causality biases are closely tied to the literal meanings of verbs, that could explain their relative robustness in the face of manipulations of non-linguistic world knowledge, except where the task strongly invokes such knowledge (as I suggested above for the Brown and Fish causal attribution task).

Conclusion

⁵ Brown and Fish (1983b) used a more inclusive definition of experiencer-subject and experiencer-object verbs than I employed above. I follow Hartshorne and Snedeker (submitted) in excluding cognition verbs (*know, understand*), perception verbs (*see, hear*), and other non-emotion verbs. There has otherwise been relatively little debate about state verbs, with most authors adopting Brown and Fish's proposal. Action verbs have generated considerably more controversy (Au, 1986; Rudolph & Forsterling, 1997; Semin & Fiedler, 1991).

Implicit causality has garnered considerable interest because it appeared to show extremely rapid use of nonlinguistic world knowledge in language processing. The experiments above show the relevant results were misinterpreted. This should not, however, diminish the importance of implicit causality phenomena in the various subdisciplines of psychology. That different implicit causality tasks elicit different causal attributions and are more or less affected by world knowledge presents an opportunity for studying different factors underlying causal reasoning under different conditions, motivating both a careful reanalysis of the literature and the conducting of numerous future studies.

- Arnold, J. E., Eisenband, J. G., Brown-Schmidt, S., & Trueswell, J. C. (2000). The rapid use of gender information: Evidence of the time course of pronoun resolution from eyetracking. *Cognition*, *76*, B13-B26.
- Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say thee uh you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914-930.
- Au, T. K. (1986). A verb is worth a thousand words: the causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language*, *25*, 104-122.
- Baayen, R. H. (2008). *Analyzing Linguistic Data*. Cambridge, UK: Cambridge University Press.
- Bates, D. M., & Sarkar, D. (2007). lme4: Linear mixed-effects models using Eigen and syntax (Version R package version 0.9975-12).
- Brown, R., & Fish, D. (1983a). Are there universal schemas of psychological causality? *Archives de Psychologie*, *51*, 145-153.
- Brown, R., & Fish, D. (1983b). The psychological causality implicit in language. *Cognition*, *14*, 237-273.
- Brown, R., & Van Kleeck, M. H. (1989). Enough said: Three principles of explanation. *Journal of Personality and Social Psychology*, *57*, 590-604.
- Caramazza, A., Grober, E., Garvey, C., & Yates, J. (1977). Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behavior*, *16*, 601-609.

- Corrigan, R. (1988). Who dun it? The influence of actor-patient animacy and type of verb in the making of causal attributions. *Journal of Memory and Language*, 27, 447-465.
- Corrigan, R. (2001). Implicit causality in language: event participants and their interactions. *Journal of Language and Social Psychology*, 20, 285-320.
- Corrigan, R. (2002). The influence of evaluation and potency on perceivers' causal attributions. *European Journal of Social Psychology*, 32, 363-382.
- Corrigan, R. (2003). Preschoolers' and adults' attributions of who causes interpersonal events. *Infant and Child Development*, 12, 305-328.
- Corrigan, R., & Stevenson, C. (1994). Children's causal attributions to states and events described by different classes of verbs. *Cognitive Development*, 9, 235-256.
- Cozjin, R., Commandeur, E., Vonk, W., & Noordman, L. G. M. (in press). The time course of the use of implicit causality in the processing of pronouns: A visual world paradigm study. *Journal of Memory and Language*.
- Crinean, M., & Garnham, A. (2006). Implicit causality, implicit consequentiality and semantic roles. *Language and Cognitive Processes*, 21, 636-648.
- Featherstone, C. R., & Sturt, P. (2010). Because there was cause for concern: An investigation into a word-specific prediction account of the implicit-causality effect. *The Quarterly Journal of Experimental Psychology*, 63, 3-15.
- Ferstl, E. C., Garnham, A., & Manouilidou, C. (in press). Implicit causality bias in English: A corpus of 300 verbs. *Behavioral Research Methods*.

Franco, F., & Arcuri, L. (1990). Effect of semantic valence on implicit causality verbs.

The British Journal of Social Psychology, 29, 161-170.

Fukumura, K., & van Gompel, R. P. G. (2010). Choosing anaphoric expressions: Do

people take into account likelihood of reference? *Journal of Memory and Language, 62*, 52-66.

Garnham, A., Traxler, M., Oakhill, J., & Gernsbacher, M. A. (1996). The locus of

implicit causality effects in comprehension. *Journal of Memory and Language, 35*, 517-543.

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry, 5*,

459-464.

Garvey, C., Caramazza, A., & Yates, J. (1974). Factors influencing assignment of

pronoun antecedents. *Cognition, 3*(3), 227-243.

Goikoetxea, E., Pascual, G., & Acha, J. (2008). Normative study of implicit causality in

100 interpersonal verbs in Spanish. *Behavior Research Methods, Instruments, & Computers, 40*, 760-772.

Greene, S. B., & McKoon, G. (1995). Telling something we can't know: experimental

approaches to verbs exhibiting implicit causality. *Psychological Science, 6*(5), 262-270.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and

possible all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition, 116*, 42-55.

- Guerry, M., Gimenes, M., Caplan, D., & Rigalleau, F. (2006). How long does it take to find a cause? An online investigation of implicit causality in sentence production. *The Quarterly Journal of Experimental Psychology*, 2006(59).
- Hartshorne, J. K., & Snedeker, J. (submitted). What is implicit causality: World knowledge, an arbitrary feature, or an effect of semantic structure.
- Hoffman, C., & Tchir, M. A. (1990). Interpersonal verbs and dispositional adjectives: The psychology of causality embodied in language. *Journal of Personality and Social Psychology*, 58, 765-778.
- Jackendoff, R. (1990). *Semantic structures*. Cambridge, MA: The MIT Press.
- Kasof, J., & Lee, J. Y. (1993). Implicit causality as implicit salience. *Journal of Personality and Social Psychology*, 65, 877-891.
- Kehler, A. (2002). *Coherence, reference, and the theory of grammar*. Stanford, CA: CSLI Publications.
- Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, 25, 1-44.
- Koornneef, A. W., & Van Berkum, J. J. A. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, 54, 445-465.
- Kuperberg, G. R., Paczynski, M., & Ditman, T. (in press). Establishing causal coherence across sentences: An ERP study. *Journal of Cognitive Neuroscience*.
- LaFrance, M., Brownell, H., & Hahn, E. (1997). Interpersonal verbs, gender, and implicit causality. *Social Psychology Quarterly*, 60, 138-152.

- Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. Chicago, IL: University of Chicago Press.
- Lombrozo, T. (2010). Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions. *Cognitive Psychology, 61*, 303-332.
- Long, D. L., & De Ley, L. (2000). Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language, 42*, 545-570.
- Maas, A., Salvi, D., Arcuri, L., & Semin, G. (1989). Language use in intergroup contexts: the linguistic intergroup bias. *Journal of Personality and Social Psychology, 57*(6), 981-993.
- Mannetti, L., & De Grada, E. (1991). Interpersonal verbs: implicit causality of action verbs and contextual factors. *European Journal of Social Psychology, 21*, 429-443.
- McDonald, J. L., & MacWhinney, B. (1995). The time course of anaphor resolution: Effects of implicit verb causality and gender. *Journal of Memory and Language, 34*, 543-566.
- McKoon, G., Greene, S. B., & Ratcliff, R. (1993). Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 1040-1052.
- Pesetsky, D. (1995). *Zero Syntax: Experiencers and Cascades*. Cambridge, MA: The MIT Press.
- Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language & Cognitive Processes, 22*, 780-788.

- Pyykkonen, P., & Jarvikivi, J. (2010). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology, 57*, 5-16.
- R-development-core-team (2005). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, <http://www.R-project.org>.
- Rudolph, U. (2008). Covariation, causality, and language: Developing a causal structure of the social world. *Social Psychology, 39*, 174-181.
- Rudolph, U., & Forsterling, F. (1997). The psychological causality implicit in verbs: a review. *Psychological Bulletin, 121*(2), 192-218.
- Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition, 71*, 109-147.
- Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and languages. *Journal of Personality and Social Psychology, 54*, 588-568.
- Semin, G. R., & Fiedler, K. (1991). The linguistic category model, its bases, applications and range *European Review of Social Psychology* (pp. 1-30). Chichester, England: Wiley.
- Semin, G. R., & Marsman, J. G. (1994). "Multiple inference-inviting properties" of interpersonal verbs: Event instigation, dispositional inference, and implicit causality. *Journal of Personality and Social Psychology, 67*, 836-849.

Stewart, A. J., Pickering, M. J., & Sanford, A. J. (2000). The time course of the influence of implicit causality information: Focusing versus integration accounts.

Journal of Memory and Language, 42, 423-443.

Van Berkum, J. J. A., Koornneef, A. W., Otten, M., & Nieuwland, M. S. (2007).

Establishing reference in language comprehension: An electrophysiological perspective. *Brain Research*, 1146, 158-171.

Whorf, B. L. (Ed.). (1956). *Language, Thought and Reality: Selected Writings of*

Benjamin Lee Whorf. Cambridge, MA: The MIT Press.

APPENDIX: Verbs

Causal verbs: weakened, balanced, softened, strengthened, improved, revived

Experiencer-object verbs: affected, puzzled, satisfied, frustrated, bored, aroused

Experiencer-subject verbs: resented, hated, despised, disliked, admired, respected

Judgment verbs: blamed, denounced, cursed, condemned, excused, criticized