I heard he might be guilty, so he probably is: Uncertain evidence statements in iterative reproductions of crime stories

Anonymous CogSci submission

Abstract

[jd: to do]

Keywords: iterated narration; transmission chains; crime stories; suspect; guilt

[ek: General notes: make up your mind about generations vs. reproduction; original stories vs. seeds; stories vs. story-type vs. condition,...]

Introduction

One of the central goals of language use is the exchange of information. New information is obtained by reading the newspaper, listening to a friend, etc., and often immediately communicated as stories to other people it may be relevant to. Yet this process of iterated reproduction is not innocuous: the original story may be distorted or altered by various sources of noise, including cognitive biases, memory reconstruction processes, or other limits on information processing capacity (Bartlett, 1932; ?, ?; Hills, in press; Mesoudi & Whiten, 2004). The game of Telephone is essentially a caricature of this process: the first person whispers a sentence to their neighbor, who in turn passes it on to the next person, and so on. The last person in the transmission chain announces the sentence they ended up with, which often differs remarkably from the initial seed story. This simple game nicely exemplifies the information loss and distortion that is associated with repeated exposure and reproduction of information.

Bartlett (1932) first introduced the methodology of transmission chains, i.e., chains of story reproductions, as a scientific method. In a series of transmission chain studies, using stories such as Native American tales or sport reports for reproduction, he observed a significant information loss in the stories over generations of reproductions. He also reported that the content of the reproduced stories increasingly aligned with the reproducing author's prior beliefs (for a more recent demonstration, see ?, ?). Bartlett used these observations as a foundation for his theory of memory retrieval involving reconstruction processes.

In recent years, the transmission chain method has undergone a revival in cognitive and social psychology. Mesoudi and Whiten (2004) showed that with each iteration, described events become more abstract. Further research showed that [ek: gender stereotypes: Bangerter 2000, Kashima 2000; cognitive biases: Kalish 2007, Griffiths 2007/2008; Stubbersfield 2015/2017; Hills/Jagiello 2018]. In linguistics, the transmission chain method has been used to study questions related to the acquisition and regularization of grammars and lexicons using iterated language learning paradigms (?, ?, ?, ?, ?, ?, ?, ?, ?, ?).

The transmission chain method thus presents an exciting opportunity for asking questions at the interface of linguistics and psychology. In particular, while previous studies have focused particularly on XXX, we use the method here to assess how the content of retold stories influences the interpretation of various aspects of those stories, by collecting judgments for each story independently collected judgments of aspects of the stories as a first step towards analyzing how the content of reproduced stories changes as a function of construct a corpus of iteratively retold stories, which we subsequently collect judgments from

Experiment 1: corpus collection

[ek: ...]

Methods

74 Stanford students participated in this online study for course credit. We constructed five stories (*seeds*) that marked the beginning of each reproduction chain. Stories were written in the style of short news articles and followed a similar structure. They reported a crime or moral rule violation that occurred, the authorities' determination of and search for the perpetrator(s), and the possible punishment the suspect(s) would face if found guilty. Furthermore, each of these five seed stories occurred in one of two conditions: a *weak evidence* and a *strong evidence* condition. Evidence strength was manipulated in the final sentence of the story (see example seed in Table 1).

Each participant read and reproduced five stories. For each story, they were either assigned to read and reproduce the seed story or continue an already started reproduction chain where they read and reproduced a reproduction from previous participants. The assignment was random. On each trial,

Table 1: Example of a seed story used in Exp. 1.

<table>
<tr><td colspan="2">In late December 2017, a couple in Iowa was checking on their 50 beehives when they discovered a tragic scene. The hives had been overturned and hacked apart, and the equipment had been thrown out of the shed and smashed. This destruction caused the death of about half a million bees and approximately $60,000 in property damage. Nearly three weeks later, police arrested two boys (12 and 13 years old) who, allegedly, were responsible for the damage. The charges against them include criminal mischief, burglary, and offenses to an agricultural animal facility. Since they are still minors, they will be charged in juvenile court where they face up to 10 years in prison and fines of up to $10,000 if convicted.</td></tr>
<tr><td>(<i>strong evidence condition</i>)
Police officials explained that the investigation is still in progress, but the evidence so far overwhelmingly speaks to the guilt of the suspects.</td><td>(<i>weak evidence condition</i>)
Police officials explained that the investigation is still in progress, and the evidence so far doesn't warrant rushed conclusions about the guilt of the suspects.</td></tr>
</table>

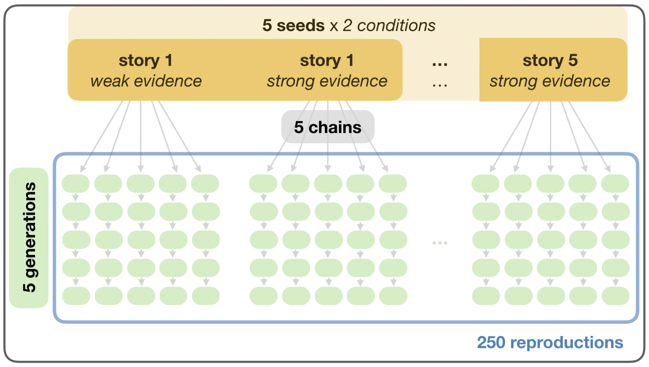

Figure 1: Overview of corpus of stories collected in Exp. 1.

participants first read a story. They were told to click the 'Continue' button when they were confident that they had internalized the story. Once they clicked the button, the story disappeared and they were asked to reproduce it freely in a text field. Order of stories was randomized.

Results

Participants produced 370 stories. For each seed, we defined a complete chain as one that has 5 reproductions/generations. For subsequent analysis, we randomly selected 50 complete chains, evenly distributed across stories and conditions. This yielded a corpus of 250 reproductions (5 seeds in 2 conditions with 5 complete chains each, see Figure 1). This was the maximal set of complete chains that was present in every condition for each seed. This corpus is a rich source of linguistic information which merits detailed investigation. Yet, with an eye to clear operationalizability, we focus here on a few general features, which we will subsequently use as predictors in the analyses of Exp. 2 below.

Story length. As shown in Figure 2, the number of words of a reproduction decreased across generations ($\beta = -17.12$, $SE = 1.02$, $t = -16.79$, $p < 0.0001$), replicating a well-known phenomenon in reproduction studies (Bartlett, 1932). While the original generation 0 seeds consisted on average of 159 words, that number dropped to 25 by generation 5.

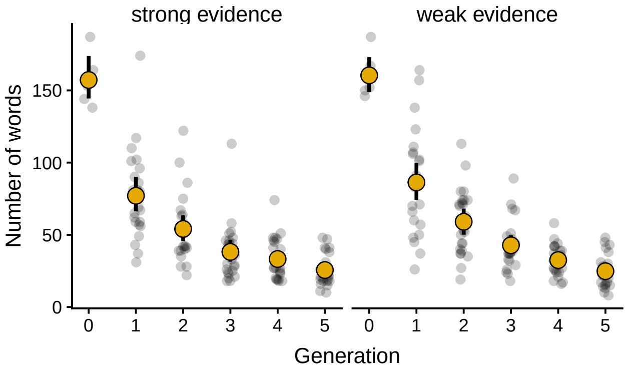

Figure 2: Mean story length in number of words by generation in strong (left) and weak (right) evidence condition. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories.

Examples of reproductions of the seed in Table 1 from generation 1 and 5 are shown in (1) and (2) below. There is no significant effect of condition.

(1) In late December 2017, a couple in Iowa went to check on their beehives. They found a tragic scene: their hives had been overturned and their equipment and facilities had been ransacked. A few weeks later, the police arrested a 12-y.o. and 13-y.o. for the crime. They are charged with multiple offenses, with fines up to $100,000 and up to 10 years in prison, yet will be tried as minors. The trial hasn't happened yet, but they seem guilty.

(2) A 12 and 13 year old were arrested for destroying a beehive, and face up to 10 years of jail time.

Similarity of seeds and reproductions. To assess the similarity of reproductions and their seed stories quantitatively, we computed the Jaccard distance between each reproduction and its generation 0 seed. Jaccard distance captures the amount of overlap between two stories in the following way:

$$D_J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

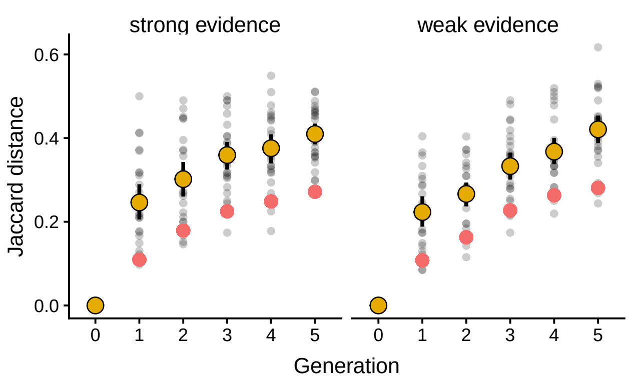

Figure 3: Mean Jaccard distance between seed and reproductions by generation in strong (left) and weak (right) evidence condition. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories, red dots indicate the lowest possible distance given the mean length of the stories.

where X is the set of words in the reproduction and Y the set of words in the respective original seed story. In this case, we took words as the basic unit over which distance was computed. Figure 3 shows that D_J increased across generations ($\beta = 0.05$, $SE = 0.00$, $t = 14.17$, $p < 0.0001$). There is no significant effect of condition. This is not surprising given that as story length decreases, D_J between seed and any of its reproductions necessarily increases. However, D_J increased more strongly than expected if the difference between stories was only due to the decrease in length [ek: is this true?] [mf: this is tricky. I wouldn't know how to *quickly* check this; first thing that comes to mind is, unsurprisingly, simulation; we could sample random subsets of words corresponding to the mean story lengths and check JD for that; but that's time consuming and I don't think that this is worthwhile currently; I'd rather suggest we formulate this differently, weaker, less prone to criticism if we do not back it up], suggesting that information was lost across generations. This can also be observed qualitatively in the comparison of the representative examples (1) and (2) above.

Proportion of Hedges. As a proxy for vagueness, we extract the number of hedges per story relative to its length. The seed stories were designed to contain various hedges, such as "nearly", "about", "up to" or "allegedly". Figure 4 shows the proportion of hedges that remain throughout the generations. We can see that this proportion decreases over generations ($\beta = -0.01$, $SE = 0.00$, $t = -4.16$, $p < 0.0001$). The reproductions therefore appear to get less mitigated over iterations. There is no significant effect of condition.

In summary, our investigation of the corpus so far revealed that the length of the stories, the similarity to the seed story and the proportion of hedges decreases over generations.

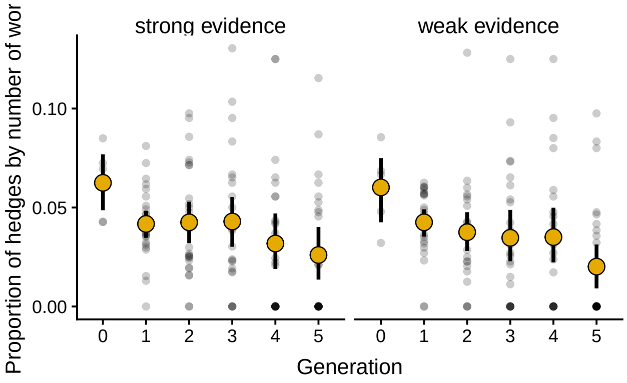

Figure 4: Mean proportion of hedges (by number of words) in strong (left) and weak (right) evidence condition. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories.

Experiment 2: story ratings

In order to assess the extent to which, as a function of the originally provided evidence, the generation of reproduction affects readers' perception of various aspects of the stories we collected judgments from a second group of independent readers of the reproductions. We are particularly interested in aspects related to the uncertainty of presented evidence, and so collected intuitive judgments related to the suspect's perceived guilt and the strength of evidence but also concerning the readers' general attitude towards the author and the story.

Methods

5392 participants were recruited over Amazon Mechanical Turk. Each participant read one story from the 250 story corpus reported in the previous section, and answered twelve questions about the story (including four attention checks). They indicated their response by moving a slider on a continuous scale. Each question was shown in isolation in a randomized order. Participants spent on average two to three minutes on this experiment and were paid $0.60 ($12-$18 per hour). The story was visible throughout the experiment.

The list of questions asked is provided in (3) to (10). Questions (3) - (7) assessed the extent to which the reader believes the suspect(s) is/are guilty of the alleged crime. Questions (8) - (10) assessed the reader's trust in the author, the extent to which they considered the story to be objectively written, and the extent to which they felt emotionally connected to the story. Overall, participants were asked eight questions of interest and four attention check questions designed to filter out participants who were just clicking through the experiment. The attention checks were counter balanced and asked about the likelihood that the passage is a Greek fairy tale, a Bible quote, contains more than five words, and that the story involves *X*, where X was replaced by a topic which was likely to occur in the story (e.g., *bees* for the story in Table 1)

(3) How strong is the evidence for the suspect's / suspects'

guilt?

(4) How likely is it that the suspect is / the suspects in the crime are guilty?

(5) How likely is a conviction of the suspect(s) in the crime?

(6) How justified would a conviction of the suspect(s) in the crime be?

(7) How much does the author believe that the suspect is guilty?

(8) How much do you trust the author?

(9) How objectively / subjectively written is the story?

(10) How affected do you feel by the story?

Results

We excluded 12 participants because they completed the study multiple times and another 535 because they failed at least two of the attention check questions. This leaves us with 4573 participants (84.8% of the original set). After exclusions, each reproduction received on average 17 ratings, ranging from 9 to 22 and two outliers with 27 and 38 ratings[1]. The original seed stories received between 25 and 31 ratings.

Mean slider ratings are shown in Figure 5. Qualitatively, there is a strong effect of condition in the guilt related measures, i.e., *strength of evidence*, *suspect guilt*, *suspect conviction*, *suspect conviction justified* and *author's belief in guilt*. In the responses to these measures, the ratings for the strong evidence condition are declining over generations, but consistently higher than for the weak evidence condition. In other words, the different guilt measures retain higher guilt judgments than the weak condition over all generations. However, the judgments of the weak conditions pattern differently. In contrast to a decline in the ratings over generations, the judgments in the weak condition stay constant or increase. This narrows the distance between the judgments in the strong and weak conditions. [ek: In sum,] the stories for the two different conditions become more similar over generations with respect to the guilt measures.

The difference between the conditions is far smaller (if existent at all) in the responses to the reader's *trust in the author*, the *subjectivity of the story* and the *reader's emotional engagement*. *Trust in the author* and the *reader's emotional engagement* decline over generations of reproductions, independent of the condition. Interestingly, the measure on how subjectively the story is written does not change over generations and remains on the "objective" side of the scale.

[ek: maybe make a note that there is no sign of convergence to .5]

The judgments were analyzed using linear mixed effects models. For each question, slider rating was predicted from fixed effects of generation (reference level: 0), condition (reference level: strong), and their interaction. The model also included random by-story intercepts. An overview of the results is shown in Table 2. Each row presents the model results for one of the questions and the columns show the model outcomes for fixed effects.

In the qualitative analysis, we have observed a strong condition effect in the guilt measures. This is reflected in the model results. In the non-guilt measures, story subjectivity shows a significant effect on condition. The trust in the author and the reader's emotional engagement appear to be independent of the condition and therefore cannot explain the difference that we see in the guilt related measures. Even though the subjectivity shows an effect of condition, it is too small [ek: rephrase this...] to explain these effects in the guilt measures either. [ek: We also know that the length and Jaccard distance (and maybe hedges?) doesn't differ in the two conditions, so...] This suggests that the strong effects of condition in the guilt measures [ek: are about some more subtle changes in the actual language and content of the stories].

With respect to the guilt measures, the quantitative analysis suggests clear effects of generation in the strength of the evidence, the likelihood of conviction, and whether the author believes in the suspect's guilt. Those three guilt measures show a very similar pattern in their main effect structure. A simple effects analysis further supports this grouping intuition, since none of them show an effect in the interaction between *weak* evidence and generation, but all of them show an effect in the interaction between *strong* evidence and generation. This group stands in contrast to the other two guilt measures, asking about the suspect's guilt and whether a conviction would be justified. These two do not show a main effect in generation and only a marginal effect in the simple effects analysis for the interaction between the strong evidence condition and generation.

Trust in author and the *reader's emotional engagement* did not show a main effect of condition, but of generation. The further it gets from the original seed, the less trust the readers have in the author and the less emotional engagement the readers feel.

In this second part of the analysis, we focus on the measures of *suspect guilt* and *author's belief in guilt*. In the first case, we ask for the participant's personal judgments on whether the suspect is guilty, whereas in the second case we ask about the participant's belief about what the author of the story believes. We will now analyze the differences between these two measures by specifically investigating the impact of length, Jaccard distance and proportion of hedges [ek: see results in 1].

In Table 2, we present the statistical results computed with linear fixed effects of generation, condition and their interaction. For *suspect guilt* and *author's belief in guilt* we added fixed effects of one of the linguistic metrics (length, Jaccard distance, proportion of hedges) and its interaction with condition. Since we expect a strong correlation of the lin-

[1] The outliers are due to a mistake in the recruitment process.

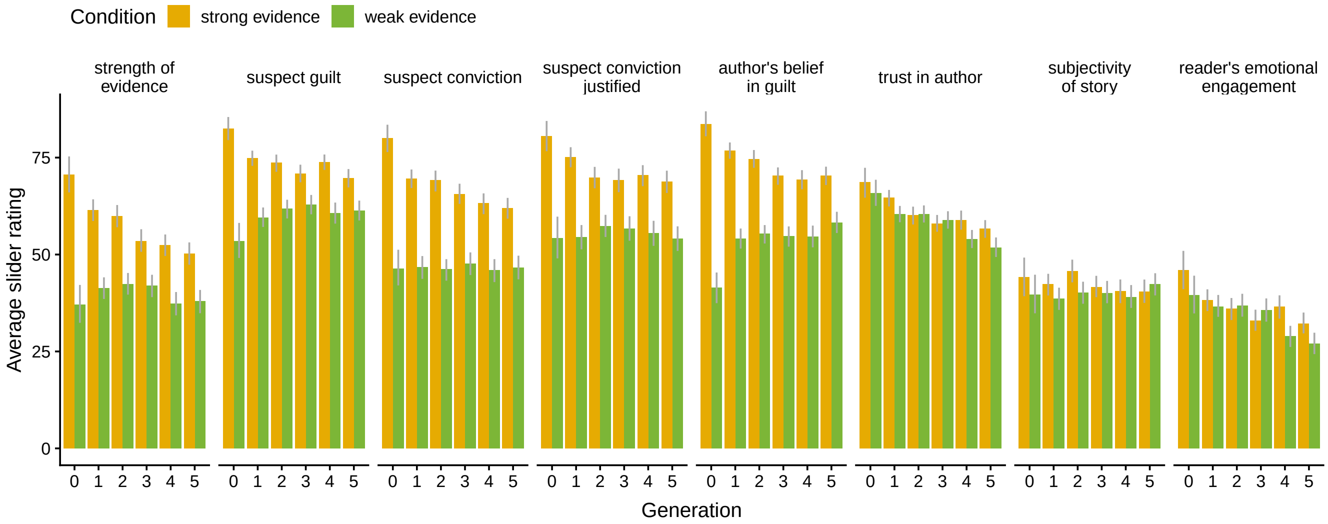

Figure 5: Mean ratings in strong (orange) and weak (green) evidence condition for each dimension (facets).

Table 2: Model output for each fixed effect (condition, generation, and their interaction) for each rated question (rows).

	condition			generation			condition*generation		
	β	SE	p	β	SE	p	β	SE	p
strength of evidence	-23.25	4.09	<0.0001***	-3.42	0.89	<0.001***	2.59	1.26	<0.05*
suspect guilt	-17.28	3.40	<0.0001***	-1.34	0.74	<0.08	1.90	1.05	<0.08
suspect conviction	-27.01	4.15	<0.0001***	-2.79	0.90	<0.01**	2.74	1.28	<0.05*
suspect conviction justified	-19.02	4.35	<0.0001***	-1.69	0.95	<0.08	1.43	1.34	<0.29
author's belief in guilt	-27.53	3.72	<0.0001***	-2.14	0.81	<0.01**	3.42	1.15	<0.01**
trust in author	-0.82	2.25	<0.72	-1.94	0.49	<0.001***	-0.54	0.70	<0.44
subjectivity of story	-6.12	2.21	<0.01**	-0.86	0.49	<0.08	1.40	0.69	<0.05*
reader's emotional engagement	0.85	2.99	<0.78	-1.49	0.65	<0.05*	-1.11	0.92	<0.24

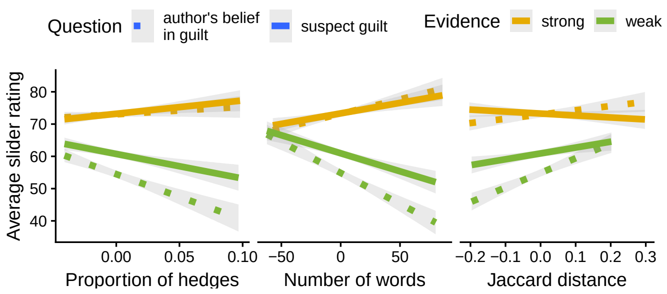

Figure 6: .

guistic metrics with generation, we take the residuals of the generations and the linguistic metrics. A model comparison was conducted using a Chi-squared ANOVA test. The model comparison revealed that a model which includes any of the linguistic metrics predicts *author's belief in guilt* significantly better than the baseline [ek: include stats]. This holds similarly for *suspect guilt* when including the Jaccard distance and length, but not the proportion of hedges [ek: include stats]. Figure (6) visualizes the differences between questions and conditions for each of the linguistic metrics. Simple effects analysis revealed that none of the linguistic metrics has an effect in the strong condition. However, in the weak evidence condition an increase in proportion of hedges and number of words significantly decreases the guilt rating overall. In the same condition, increasing Jaccard distance results in higher guilt ratings.

Even though the question about *suspect guilt* and *author's belief in guilt* receive very similar ratings in the strong conditions, the weak condition shows a clear distinction. Overall participants judge the suspect's guilt in the weak condition higher then what they consider the author's belief about the suspect's guilt. Furthermore, this difference in judgment decreases the smaller the proportion of hedges, the shorter the stories and the bigger the Jaccard distance. This suggests that the inclusion of hedges rather informs our beliefs about the author's beliefs and less about the suspect's guilt. Analogously, the bigger the Jaccard distance to the seed story, the more similar the guilt judgments become. The observations for Jaccard distance in the weak evidence condition suggest that the stories change in the direction to accommodate a reader's beliefs about the suspect's guilt. [ek: this supports the convergence to prior.]

Conclusion

[ek: insert intro sentence]

First we constructed a corpus which consists of 10 original seed stories and 250 reproductions. Half of the seed stories suggested strong and the other half weak evidence that speaks for the suspect's guilt. Our corpus collection replicates previously found effects of a decrement in length over generations of reproductions ((Bartlett, 1932)). Furthermore, the stories become less similar to the original seed story and the proportional number of hedges decreases.

In a second study, we obtained subjective ratings for each story regarding the question of guilt, conviction, the author of the story, the subjectivity of the story and the reader's emotional engagement. Our results suggest **that, for one, the subtle manipulation of varying evidential strength in the original seed stories did have a lasting effect on reproductions lasting across several generations.** We also find effects of generation. It is plausible that trust in the author and the reader's emotional engagement decrease because the reproductions become shorter with each iteration. It is less likely that this explanation holds for the generation effects in *strength of evidence*, likelihood of *suspect conviction* and the *author's belief in the suspect's guilt*. In the ratings for these questions we see a stronger difference between conditions that the changes in length, similarity and the proportion of hedges could not account for. [ek: Therefore, it must be something about the content of the stories itself that changes.]

Most interestingly, it seems that readers became, over generations, more inclined to endorse the guilt of the suspect(s) in the weak evidence condition, despite no similar increase in ratings of the strength of evidence. This may be due to a baseline or prior level of conviction that a suspect is guilty. And it would suggest that, indeed, subtle evidential information and uncertainty about evidence is washed out over several linguistic transmissions.

[mf: I have tried to give this some top-level interpretation, but I have to admit that I'm pretty lost when it comes to interpreting the results.]

[ek: discuss differences between stories with in- and out-group effects for smuggler and professor] [ek: next steps?]

[mf: it is not common to have a discussion section after the conclusions. let's put it all in one section]

References

Bartlett, F. C. (1932). Remembering: An experimental and social study. *Cambridge: Cambridge University*.

Hills, T. T. (in press). *The dark side of information proliferation.* (Journal: Perspectives on Psychological Science)

Mesoudi, A., & Whiten, A. (2004). The hierarchical transformation of event knowledge in human cultural transmission. *Journal of Cognition and Culture*, *4*(1), 1–24.

Van Prooijen, J.-W. (2006). Retributive reactions to suspected offenders: The importance of social categorizations and guilt probability. *Personality and Social Psychology Bulletin*, *32*(6), 715–726.

	condition			generation			condition*generation			simple effects		
	β	SE	p	β	SE	p	β	SE	p	weak	str*gen	we*g
evidence	-23.25	4.09	<0.0001***	-3.42	0.89	<0.001***	2.59	1.26	<0.05*	***	***	
suspect guilt	-17.28	3.40	<0.0001***	-1.34	0.74	<0.08	1.90	1.05	<0.08	***	.	
suspect conviction	-27.01	4.15	<0.0001***	-2.79	0.90	<0.01**	2.74	1.28	<0.05*	***	**	
conviction justified	-19.02	4.35	<0.0001***	-1.69	0.95	<0.08	1.43	1.34	<0.29	***	.	
author's belief in guilt	-27.53	3.72	<0.0001***	-2.14	0.81	<0.01**	3.42	1.15	<0.01**	***	**	
trust in author	-0.82	2.25	<0.72	-1.94	0.49	<0.001***	-0.54	0.70	<0.44		***	***
subjectivity of story	-6.12	2.21	<0.01**	-0.86	0.49	<0.08	1.40	0.69	<0.05*	**	.	
reader's emotion	0.85	2.99	<0.78	-1.49	0.65	<0.05*	-1.11	0.92	<0.24	*	***	

Table 3: Model output for each fixed effect (condition, generation, and their interaction) for each rated question (rows). [jd: simple effects results should not be reported in this table – this is just here for us, right?][ek: yes]