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

Anonymous CogSci submission

Abstract

Transmission of information by means of language is a potentially lossy process. Especially adjunct information, such as the graded degree of evidence, is *prima facie* a piece of information that seems likely to be distorted by reproduction noise. To investigate this issue, we present the results of a two-step iterated narration study: first, we collected a corpus of 250 crime story reproductions that were produced in parallel reproduction chains of 5 generations in depth, for 5 different seed stories; a second separate large-scale experiment then targeted readers' interpretation of these reproductions. Crucially, strength of evidence for the guilt of each story's suspect(s) was manipulated in the initial seed stories. Across generations, readers' guilt perceptions decreased when the evidence was originally strong, but remained stable when evidence was originally weak. Analysis of linguistic measures revealed that dissimilarity between a seed story and its reproduction, story length, and amount of hedging language affected the readers' own guilt perception and the readers' attribution of guilt perception to the author differently. The results provide evidence that evidential information indeed influences guilt perception in complex ways.

Keywords: experimental pragmatics; iterated narration; transmission chains; uncertain evidence

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; Mesoudi & Whiten, 2004; Griffiths & Kalish, 2007; Hills, in press). 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. 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 descriptions of everyday events, such as visits to a restaurant, became more abstract, in line with hierarchically organized script knowledge. Other research showed that reproductions can be influenced by cultural, racial and gender stereotypes (e.g., Kashima, 2000). The iterated transmission method has therefore also been used as a tool to investigate cognitive biases in general (e.g., Kalish, Griffiths, & Lewandowsky, 2007). In evolutionary linguistics, the transmission chain method has been used to study experimentally how iterated learning of a language exerts a selective pressure on language itself, so that learning biases create an indirect pressure on languages to be efficiently learnable (e.g., Scott-Phillips & Kirby, 2010; Kirby, Griffith, & Smith, 2014).

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 properties of the reproductions themselves, we here present an extension in which we investigate an external readership's interpretative perspective on the reproduced texts. We achieve this by a second experiment that uses as materials the output from the previous iterated transmission experiment. The stories used as seeds are five crime or ethical violation stories based on true events (animal smuggling, arson, sexual assault, beehive destruction, and email scams). Each seed started out with both a weak and a strong evidence version (see Table 1). This manipulation has successfully been used by (Van Prooijen, 2006) to uncover in- and out-group effects in guilt judgments of suspects. Similarly to that study, the different conditions were implemented by adding a last sentence to each story that either suggested strong or weak evidence for the suspect's guilt. To evaluate the interpretations of the stories that readers arrive at, we collected answers to eight questions regarding the readers' perception of the stories' suspect(s), the readers guilt perception attributed to the author of the story, as well as, somewhat less importantly, indexes of more general author and reader related features, such as trustworthiness and subjective engagement.

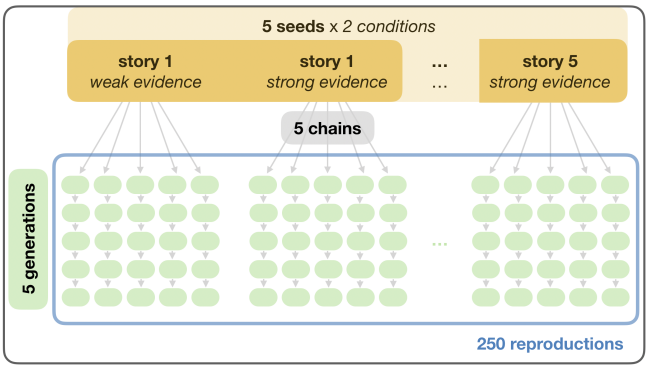

Figure 1: Overview of corpus of stories collected in Exp. 1.

Experiment 1: corpus collection

Methods

74 undergraduate 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, 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.

Proportion of hedges. As a proxy for vagueness, we extracted 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". As shown in Figure 2, the proportion of hedges decreased in each generation ($\beta = -0.01$, $SE = 0.00$, $t = -4.16$, $p < 0.0001$), suggesting that participants portrayed the stories with more certain language over generations. There was no significant effect of evidence condition on proportion of hedges.

Story length. As shown in Figure 2, the number of words in a story 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. Examples of reproductions of the seed in Table 1 (strong condition) from generation 1 and 5 are shown in (1) and (2) below. There was no significant effect of evidence condition on story length.

(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 between seed stories and their reproductions quantitatively, we computed the Jaccard distance between each reproduction and its generation 0 seed. Jaccard distance ranges between 0 and 1 (where 1 indicates greatest distance) and captures the amount of overlap between two stories in the following way:

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

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 2 shows that D_J increased across generations ($\beta = 0.05$, $SE = 0.00$, $t = 14.17$, $p < 0.0001$). This is not surprising given that as story length decreases, D_J between seed and any of its reproductions necessarily increases. However, we will see later that length and D_J have different effects on story interpretation. There was no significant effect of evidence condition on Jaccard distance.

In sum, in a corpus of 250 reproductions of 5 seed stories, the length of the stories, the similarity to the seed story, and the proportion of hedges decreases over generations, regardless of the initial evidence strength condition.

Experiment 2: story ratings

In order to assess the extent to which, as a function of the originally provided evidence, the generation of reproduction

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

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.	
(*strong evidence condition*) Police officials explained that the investigation is still in progress, but the evidence so far overwhelmingly speaks to the guilt of the suspects.	(*weak evidence condition*) 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.

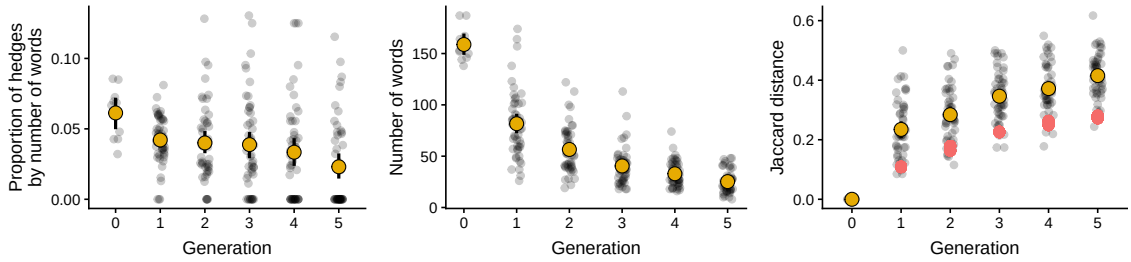

Figure 2: Mean of the three linguistic metrics (proportion of hedges (by number of words), number of words and Jaccard distance) over generations of reproductions. Error bars indicate bootstrapped 95% CIs. Orange dots indicate generation mean, gray dots are individual stories. The red dots indicate the lowest possible distance given the mean length of the stories.

affects readers' interpretation of various features of the stories we collected judgments from a second group of independent participants. We were particularly interested in features related to the uncertainty of presented evidence and the associated judgments of suspect guilt. We also collected judgments 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 (slider endpoints were underlyingly coded as 0 - 100). 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.

(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

Exclusions. 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 left 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. The original seed stories received between 25 and 31 ratings.

Analysis procedure. Of main interest was whether participants give judgments of suspect guilt in line with the originally provided evidence, whether those judgments change

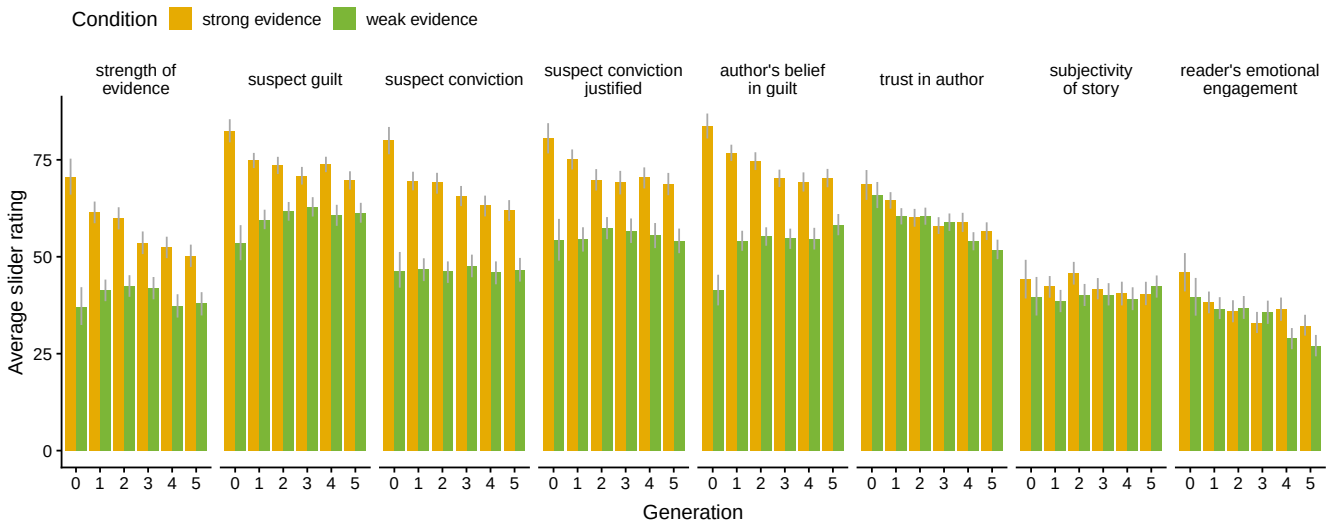

Figure 3: Mean ratings in strong (orange) and weak (green) evidence condition for each dimension (facets).

over generations, and whether those judgments pattern with related measures of evidence for guilt, probability of conviction, justification of conviction, and attributed suspect guilt (i.e., an estimate of the author's belief in the suspect's guilt). We refer to these measures as *guilt related measures*. Additionally, we analyzed trust in the author, story subjectivity, and emotional engagement as measures of secondary interest. Mean slider ratings corresponding to the analyses are shown in Figure 3. Judgments were analyzed using linear mixed effects models. For each question, slider rating was predicted from fixed effects of generation, condition (reference level: strong), and their interaction. The models 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.

Generation and evidence strength effects on guilt related measures. We observed main effects of condition on all guilt related measures (see first 5 panels of Figure 3), such that ratings in the strong condition were higher than ratings in the weak condition, suggesting that participants in Exp. 1 were sensitive to the evidence strength manipulation (reproducing stories in such a way that evidence strength information was maintained); and also suggesting that participants in Exp. 2 were sensitive to the reproduced evidence strength information in their judgments. We also observed significant or marginally significant interactions between condition and generation for all guilt related measures, such that ratings decreased across generations in the strong condition but remained stable in the weak condition.

Generation and evidence strength effects on secondary measures. The secondary measures look very different from the guilt related measures. In particular, there were no significant effects of evidence strength condition on any of the measures with the following exception: stories were rated as less subjective in the weak evidence condition in earlier generations, though subjectivity ratings did not vary as a function of generation and remained on the 'objective' side of the scale throughout. In contrast, both trust in the author and readers' emotional engagement with the story decreased across generations. This is presumably the result of the stories becoming shorter over generations (see Exp. 1) and readers therefore having less material to be emotionally affected by, and less material to build trust in the author on in later generation stories.

Preliminary discussion. It seems *prima facie* plausible that trust in the author, the subjective quality of the story, or the reader's emotional engagement with the story are important factors in readers' assessment of the described suspects' guilt. But the presented data suggest otherwise. The guilt related measures are entirely uncorrelated with the secondary measures. The evidence strength effect was expected, given the strong manipulation in the final sentence of the seed story. However, what is it that changes over generations that affects the guilt related measures in the strong and weak evidence conditions differently? What changes is the content of the stories. We next report a second set of analyses in which we assess the extent to which the linguistic features reported in Exp. 1 predict ratings in Exp. 2, focusing on the readers' assessment of suspect guilt and of attributed suspect guilt.

Effects of linguistic features on suspect guilt and author belief in suspect guilt. In this part of the analysis, we focus on the measures of *suspect guilt* and *author's belief in guilt* (attributed suspect guilt). These measures are interesting to examine in more detail because a) suspect guilt is the main issue raised in the 5 seed stories, so it is relevant to understand the linguistic conditions that lead to changes in perceived guilt; and b) while there is no obvious reason why

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

readers' ultimate beliefs and the beliefs they ascribe to the author *should* differ after reading these stories, Degen et al. (2019) showed that listeners maintain uncertainty about the state of the world even when they ascribe a strong belief to speakers. In the following, we therefore analyze for both measures the effect of the proportion of hedges in a story, story length, and dissimilarity between a story and its seed.

Results are shown in Figure 4. In order to analyze the effects of proportion of hedges, story length, and Jaccard distance on the two guilt measures of interest, we asked whether the linguistic features explained variance above and beyond generation. To assess this, we first residualized each feature against generation, due to the substantial correlations between the features and generation observed in Exp. 1. The final mixed effects linear regression models predicted slider rating for each of the two measures and each of the three linguistic features of interest from main effects of evidence strength condition, residualized linguistic feature, generation, the interaction between evidence strength condition and generation, and the interaction between evidence strength condition and residualized linguistic feature.[1]

We observed significant interactions between evidence strength condition and generation-residualized linguistic feature for two of the three linguistic features for both attributed suspect guilt (hedge proportion: $\beta = -162.93$, $SE = 58.98$, $t = -2.76$, $p < .01$, story length: $\beta = -0.30$, $SE = .06$, $t = -4.38$, $p < .0001$, Jaccard distance: $\beta = 20.71$, $SE = 17.3$, $t = 1.2$, $p < .24$) and suspect guilt (hedge proportion: $\beta = -111.54$, $SE = 54.5$, $t = -2.05$, $p < .05$, story length: $\beta = -0.18$, $SE = .06$, $t = -2.93$, $p < .01$, Jaccard distance: $\beta = 26.34$, $SE = 18.6$, $t = 1.42$, $p < .16$).

Simple effects analysis revealed no evidence of an effect of the linguistic metrics in the strong evidence condition. However, in the weak evidence condition an increase in proportion of hedges and number of words significantly decreased guilt ratings in both measures. In contrast, increasing Jaccard distance resulted in higher guilt ratings only in the weak condition.

[1]Nested model comparison revealed that the inclusion of the residualized linguistic feature fixed effects were justified in all linguistic features for author's belief in guilt and for story length in suspect guilt.

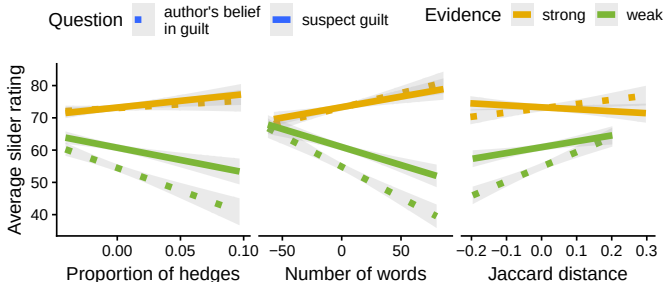

Figure 4: Linearly smoothed mean slider ratings as a function of generation-residualized proportion of hedges in story (left), number of words (middle), and Jaccard distance (right). Suspect guilt ratings shown in solid lines, author belief in suspect guilt ratings shown in dashed lines. Gray ribbons indicate 95% confidence intervals.

Even though the questions about suspect guilt and attributed guilt received very similar ratings in the strong conditions, the weak conditions showed a clear difference. Overall, participants provided higher ratings for the suspect's guilt than for attributed guilt in the weak condition. Furthermore, this difference in judgment decreased with smaller proportions of hedges, shorter stories, and (with an only trending interaction) bigger Jaccard distances. This suggests that the inclusion of hedges has a differential effect on interpretation when evidence is portrayed as strong vs. weak. With weak evidence, hedges are taken to further weaken judgments of *attributed* guilt – but participants are inclined to disregard those weaker judgments in forming their own opinion. The same is observed with longer stories.

General discussion

In this work we investigated the effects of lossy transmission on readers' interpretation of crime stories under varying initial evidential conditions in an iterated narration paradigm. First we constructed a corpus of 5 original seed stories in

2 conditions of evidential strength for a suspect's guilt, and 250 reproductions thereof. This corpus replicates previously found effects of a decrease in story 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.

We here introduced a, to our knowledge, new experimental extension of the transmission chain paradigm, where we subjected the text reproductions for the first study to a second empirical study focusing on the interpretative effect of the reproductions on a second set of independent readers. In this way, we obtained ratings for each story on 5 guilt related measures and 3 secondary measures regarding trust in the author, story subjectivity, 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 and subsequent judgments of guilt lasting across several generations. For another, manipulation of evidence did not seem to have an effect on the readers' perception of the trustworthiness of the author, the subjectivity of the story or the general engagement readers had with the story. This is partially surprising because it seems naïvely plausible that providing weaker evidence could lead to less trustworthiness and more subjectivity. However, if reproductions are convincing and present also weak evidence in the right, there is no need to assume that the author is not trustworthy or the text more subjective.

We also observed effects of generation, which we found to be attributable to the ways the reproduced stories changed over generations. The most striking result of this investigation is, perhaps, that contrary to pessimistic expectations it did not appear to be the case that repeated reproductions of stories with nuanced degrees of evidential information would have dropped these nuances, e.g., to arrive at a black-and-white picture. Instead of increasing ratings for strength of evidence in the weak evidence conditions, we rather see a decline of perceived evidential strength over generations in the strong evidence condition. Reproducers seemed to have been rather careful in their formulations, despite the observed decrease in the proportion of hedges.

We see the main achievements of this work in the contribution of an interestingly structured text corpus, with rich empirically obtained information on readers' assessments of the individual texts. This data set enables more detailed linguistic analyses in future work. As a first step in this direction, we have here considered text-based measures related to length, similarity to the original seed and the proportion of hedges. The latter is arguably most interesting from a linguistic point of view. Interestingly, we observed that, when controlling for the effects of generation, the proportion of hedges in a reproduction influenced readers' guilt perception and attributed guilt perception, but only for the weak evidence condition, in such a way that, as seems natural, higher proportions of hedges correlated with lower guilt perception and attributed guilt perception. Future work will look more closely at the more specific contribution of different types of hedges and other types of constructions that signal information about graded evidence.

References

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

Degen, J., Trotzke, A., Scontras, G., Wittenberg, E., & Goodman, N. D. (2019). Definitely, maybe: A new experimental paradigm for investigating the pragmatics of evidential devices across languages. *Journal of Pragmatics*, *140*, 33–48.

Griffiths, T. L., & Kalish, M. L. (2007). Language evolution by iterated learning with bayesian agents. *Cognitive science*, *31*(3), 441–480.

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

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*(2), 288–294.

Kashima, Y. (2000). Maintaining cultural stereotypes in the serial reproduction of narratives. *Personality and Social Psychology Bulletin*, *26*(5), 594–604.

Kirby, S., Griffith, T., & Smith, K. (2014). Iterated learning and the evolution of language. *Current Opinion in Neurobiology*, *28*, 108–114.

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

Scott-Phillips, T. C., & Kirby, S. (2010). Language evolution in the laboratory. *Trends in Cognitive Sciences*, *14*(9), 411–417.

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.