Linguistic Experience Affects Pronoun Interpretation

Jennifer E. Arnold, University of North Carolina, Chapel Hill

Iris M. Strangmann, CUNY Graduate Center, New York

Heeju Hwang, The University of Hong Kong

Sandra Zerkle, University of North Carolina, Chapel Hill

Rebecca Nappa, University of North Carolina, Chapel Hill

Corresponding author:

Jennifer Arnold

University of North Carolina, Chapel Hill

Department of Psychology

CB#3270 Davie Hall, Rm. 337B

Chapel Hill, NC 27599-3270

Abstract

We test the hypothesis that language experience influences the cognitive mechanisms used to interpret ambiguous pronouns like *he* or *she*, which require the context for interpretation. Pronoun interpretation is influenced by both the linguistic context (e.g., pronouns tend to corefer with the subject of the previous sentence) and social cues (e.g., gaze can signal the pronoun's referent). We test whether pronoun comprehension biases are related to the individual's linguistic exposure. We focus on written language experience as a metric of linguistic exposure, given that reading experience varies widely, and can be probed with the Author Recognition Task (ART). In three experiments, people with higher ART scores assigned pronouns to the grammatical subject more consistently. ART scores correlated with some skill measures, but pronoun comprehension was not explained by working memory, theory of mind, or socioeconomic status. Our results suggest that language exposure affects language comprehension at the discourse level.

Keywords: pronoun comprehension, individual differences, print exposure, discourse

How does our linguistic experience influence the mechanisms of language processing? There is no doubt that one must hear a language to learn it, so experience must be involved at some level. Yet it is unclear whether specific individual differences in experience affect the strategies or mechanisms used to process language, in particular at the discourse level. Here we address this question by investigating individual differences in the comprehension of ambiguous pronouns like *he* and *she.*

Recent evidence suggests that lexical and syntactic processing mechanisms are modulated by language experience (Farmer, Monaghan, Misyak, & Christiansen, 2011; Farmer, Fine, Misyak, & Christiansen, 2015; James, Fraundorf, Lee, & Watson, under review; Wells, Christiansen, Race, Acheson, & MacDonald, 2009) and the frequency of syntactic structures in natural language (MacDonald, 2013; Montag & MacDonald, 2008; Tanenhaus & Trueswell, 1995). These findings raise questions about whether exposure also affects the processing of larger units of text, such as anaphoric relations between pairs of utterances. Here we test this question with respect to spoken pronoun comprehension, by asking whether individual differences in print exposure correlate with spoken pronoun comprehension biases. This question is important to examine, given that many current models of pronoun comprehension suggest that people rely on calculations of probabilities, such as the probability that the speaker would refer to a particular referent (Arnold, 1998; Kehler, Kertz, Rohde, & Elman, 2008), or the probability that the speaker would use a pronoun (Kehler et al., 2008). These theories raise questions about whether pronoun comprehension is related to the comprehender's knowledge about which referential patterns are more frequent in discourse.

One well-known finding is that pronouns are assigned to linguistically salient referents. One linguistic feature that signals contextual salience is grammatical position, where characters

mentioned in grammatical subject position are considered salient and topical (Arnold, Eisenband, Brown-Schmidt, & Trueswell, 2000; Brennan, Friedman, & Pollard, 1987; Garnham, 2001; Gordon, Grosz, & Gilliom, 1993; Stevenson, Crawley, & Kleinman, 1994). For example, in *Ella ate lunch with Leona. She had a salad,* the pronoun *she* could refer to either Ella or Leona. Nevertheless, listeners tend to assume that Ella is the referent, since she was in subject position in the first sentence, and was the first mentioned (Gernsbacher & Hargreaves, 1988). This is known as the "subject bias", where subjects are generally preferred as pronoun referents. This bias is not grammatically required, in that pronouns can also be used to refer to the object (*Birdie gave Kay a cookie, and she ate it).* Instead, the subject bias represents a general tendency. Other linguistic constraints influence listeners' preferences as well, for example a parallel syntactic function bias (Chambers & Smyth, 1998; Grober, Beardsley, & Caramazza, 1978), linguistic focusing constructions (Almor, 1999; Arnold, 1998; Cowles, Walenski, & Kluender, 2007; Foraker & McElree, 2007), or prior pronominalization (Kaiser, 2011; Kameyama, 1996).

Theoretically, these linguistic patterns have been explained in terms of the conceptual status of the referent. Pronouns tend to be used for referents that are salient, or accessible in the context (e.g., Ariel, 1990; Gundel, Hedberg, & Zacharski, 1993), and they are understood more quickly if they have a focused or salient antecedent (e.g., Garnham et al., 1996; MacDonald & McWhinney, 1995). Salience is not merely a property of the discourse context, and can even be affected by nonlinguistic, social information like pointing or gazing. (Goodrich Smith & Hudson Kam, 2012; Hanna & Brennan, 2007). A finding that is key to the current study is that pronoun comprehension is influenced by gaze and pointing gestures, where people are more likely to assign pronouns to referents when the speaker had gazed or pointed at them. (Nappa &Arnold, 2012).

However, defining salience/accessibility is difficult, which has led several researchers to suggest that probability representations instead provide a concrete mechanism for characterizing information status. For example, recently mentioned entities are considered salient in the discourse, and are good candidates as pronoun referents, as in *Lester wrote a book, and he gave it to his grandchildren.* In addition, recently mentioned entities are highly likely to feature in the subsequent discourse, and therefore are **referentially predictable**. Arnold (1998; 2001; 2010) suggested that this predictability underlies the constraining effects of the linguistic context. For example, in addition to recently mentioned entities, grammatical subjects are more likely to be re-mentioned than other entities in an utterance, other things being equal (Arnold, 1998; Brown, 1993). Thus, predictability may also underlie the subject bias in pronoun comprehension. These patterns of re-mention provide evidence to language learners about the kinds of things that speakers tend to be attending to and what is likely to be important in the upcoming discourse.

A related idea emerges in several Bayesian models of reference comprehension. For example, in Kehler & Rohde's (2013) computational model (see also Kehler, Kertz, Rohde, & Elman, 2008), pronoun interpretation depends on estimating both the probability that a particular entity will be mentioned again and the probability that the speaker will use a pronoun. Their model has been used to account for contexts such as *Dorothy impressed Lucy because she…*, where people tend to associate "she" with Dorothy. They suggest that this stems from two probabilities: first, the high probability that Dorothy would be mentioned in this context, given that she is the most likely cause of the impressing event, and second, the fact that speakers tend to use pronouns when referring to entities last mentioned in subject position, i.e. Dorothy.

Similar Bayesian models are proposed by Hartshorne, O'Donnell and Tenenbaum (2015) for pronoun comprehension, and by Frank and Goodman (2012) for reference modification.

Frank and Goodman's model addresses the use of modifiers like "blue" for shapes in a display. They calculate the probability of referring to each shape (without any discourse context) by telling people that a speaker has used an unknown word to refer to one of three shapes, and asking them to bet which object it is. Thus, they use an experimental task to estimate probability of reference. Hartshorne et al.'s model is specifically about pronoun comprehension, but it calculates the probability of the message, rather than the probability of a referent per se. Hartshorne et al. focus on implicit causality sentences like *Archibald angered Bartholomew because he is reckless*, which tap into the calculation that the speaker probably meant that Archibald is reckless, because the semantics of "angered" makes that the most probable meaning. Thus, inferences about the most probable message have direct consequences on inferences about the most probable referent.

These models raise a critical question: where do referential probabilities come from? One possibility is that people use semantic knowledge to calculate the speaker's likely content, as in Hartshorne et al.'s (2015) model. An alternate (and not mutually exclusive) possibility, which is the idea behind the current study, is that comprehenders may draw on their experience with how linguistic units are used. For example, they may learn that speakers tend to continue talking about recently mentioned entities, especially subjects (Arnold, 1998, 2001, 2010). That is, discourse-level relations are systematic, and listeners may learn which patterns are most frequent. If so, we would predict that people with greater exposure to language should learn these patterns more strongly, and be more likely to access them during language processing.

Support for this hypothesis comes from the fact that effects of both frequency and linguistic experience have been found for syntactic and lexical processing. First, studies demonstrate that frequency affects language processing. Comprehenders are faster to understand

both words (Solomon & Howes, 1951) and structures (MacDonald, 1994, 2013; Tanenhaus & Trueswell, 1995) that occur more frequently. Second, recent work even suggests that adult language users continue to implicitly learn about the frequency of linguistic structures in their environment, such that recent experience changes the way new input is processed (Farmer, Monaghan, Misyak, & Christiansen, 2011; Farmer, Fine, Misyak, & Christiansen, 2015; Fine & Jaeger, 2013; Fine, Jaeger, Farmer & Qian, 2013; James, Fraundorf, Lee, & Watson, under review; Wells, Christiansen, Race, Acheson, & MacDonald, 2009). Some support for this idea comes from studies that manipulate experience within the experiment. For example, the comprehension of relative-clause constructions was facilitated when subjects were exposed to similar structures over a 3-week training period (Wells et al., 2009). Likewise, comprehension biases changed through exposure to exemplars of syntactic constructions, especially rare ones (Fine & Jaeger, 2013). Other support comes from studies that look at individual differences in print exposure – that is, reading and exposure to books – which is one source of language experience. For example, people with greater print exposure are more accurate in written syntactic processing tasks (e.g., James et al., under review), and more likely to use literate structures like passives (Montag & MacDonald, 2015).

In addition, Oakhill and Yuill (1986) found that children classified as high-skill readers had better pronoun comprehension than low-skill readers. In their study, 7-8 year old children read sentences like “Peter lent ten pence to Liz because she needed it,” which required an inference based on implicit causality judgments. Low-skilled readers made more errors, even in stories where the two characters differed in gender (and thus the pronoun was unambiguous by gender). Given that reading skill typically correlates with reading exposure (Stanovich & West, 1989), this provides initial evidence that exposure does affect comprehension accuracy.

However, their study focused entirely on sentences requiring a semantic inference to understand a pronoun (with or without a gender cue), and it is not clear whether exposure also affects sensitivity to information-status cues like the subject bias. In addition, their study critically showed that reading skill predicted performance on a reading task. Here we ask instead whether exposure affects comprehension in spoken language.

The importance of language exposure is also indirectly supported by research on working memory. Many studies have shown that individuals vary in their working memory capacity, i.e. their ability to hold information in memory while doing cognitive tasks (Baddeley, 1992). Daneman and Carpenter (1980) developed a language-specific version of this task, the sentence span task, in which subjects judge the grammaticality of sentences while holding words in memory. They demonstrated that this verbal memory span correlates with successful language comprehension. Verbal memory span also predicts performance on discourse comprehension tasks (Calvo, 2001; Linderholm, 2002), including pronoun comprehension tasks (Daneman & Carpenter, 1980; Nieuwland & van Berkum, 2006). Yet some researchers have argued that the sentence span task is not a pure indicator of memory capacity, and that it is heavily influenced by individual variation in language experience (MacDonald & Christiansen, 2002). Indeed, Farmer, Fine, Misyak and Christiansen (2015) found that Caplan and Waters' (1999) sentence span task correlated significantly with other proxy measures of reading exposure, such as the Author Recognition Task (Stanovich & West, 1989). Thus, evidence that verbal span correlates with pronoun comprehension may suggest a role for reading experience.

Potential Effects of Linguistic Experience on Pronoun Comprehension

This study is the first to examine the relation between individual differences in reading experience and spoken pronoun comprehension. A priori, there are several possible effects of

experience. At a general level, exposure may strengthen processes that are necessary to understand language in context. For example, comprehenders need to attend to the linguistic context, and activate hypotheses about the coherence relations between utterances – for example, does this sentence describe a next event, a cause of a previous event, or something else? (Kehler, 2002). These skills underlie both spoken and written language understanding, and could potentially underlie multiple aspects of discourse processing. Another possible effect could stem from the fact that written language is more decontextualized than spoken language. If listeners are frequently exposed to decontextualized language, they may learn that the clues to the speaker's meaning are contained within language itself, decreasing the tendency to look for social cues like gazing.

More specifically, experience also provides language users with evidence about the most frequent types of reference. People could potentially track language statistics about pronouns specifically (e.g., given a pronoun, what types of things is it likely to refer to?). However, most current models of reference comprehension suggest that what is most important is referential probability overall – that is, given a particular discourse context, what is the likelihood that each entity might be referred to (Arnold, 1998; Kehler & Rohde, 2013; Rohde & Kehler, 2014)? If so, comprehenders might use referential frequency to estimate referential likelihood.

Indeed, research shows that some types of reference are more likely than others. Arnold (1998) examined patterns of reference in written children's stories, and found that references (both pronouns and other forms) are more likely to refer back to something that had appeared in the subject position of the previous clause than something that had appeared in nonsubject position. This occurs because speakers tend to stick with a topic for a period of time, re-

mentioning the same referents repeatedly, and they tend to put topical information in subject position (Chafe, 1976).

This means that exposure to texts may help people learn that subjects are likely to be rementioned. At a simple level, people may learn that subject referents tend to persist in the discourse, as in *Arthur woke up. He ate breakfast, and then he went to work*. Yet more complex sentences may be needed to solidify the representation that subjects are more likely to be re-mentioned than objects or obliques. For that, individuals need to learn that sequences like (1), where the subject is re-mentioned, are more frequent than sequences like (2), where the nonsubject is re-mentioned.

(1) Wiley gave a present to Nellie. Wiley said "Happy birthday."

(2) Wiley gave a present to Nellie. Nellie said "Thank you."

Note that these sentences without pronouns are still evidence about whether the speaker is talking about the subject (as in 1) or the prepositional object (as in 2). Thus, encountering sequences like (1) supports the conclusion that reference to subjects is more frequent (Arnold, 1998, 2001, 2010), because there is more than one referent in the utterance.

There is evidence that written texts provide just this sort of evidence. In a reanalysis of the data from Arnold (1998), Arnold, Strangmann, Hwang, & Zerkle (2018) examined only those clauses that contained more than one reference, and tallied the frequency with which subjects and nonsubjects were mentioned in the next clause. Subjects were re-mentioned 37% of the time, while objects/obliques were only mentioned again 20% of the time. This reinforces the view that subjects are a prominent, topical position.

Note that subjects are also more likely to be mentioned in spoken language, so written language is not the only source of evidence about referential patterns. Nevertheless, written

language may provide an ideal sort of input. Learning these patterns may be supported through exposure to discourses that are thematically organized, such as written narratives. Attention to these patterns may also be supported by the decontextualized nature of written language, in which social cues like gestures and gazes are not relevant.

Moreover, there is evidence that written language exposure influences spoken language processing in other domains. People with higher print exposure produce passive structures more frequently in spoken language (Montag & MacDonald, 2015). In addition, well-read speakers use discourse styles that are conventionalizations of both spoken and written strategies (Chafe & Tannen, 1987; Tannen, 1979, 1980, 1981). There is also evidence for the reverse effect, where early classroom spoken discourse supports later literacy (Cook-Gumperz & Gumperz, 1981).

If reference processing is related to reading experience, there are two broad directions the effect could take. One possibility is that reading is important for exposure to rare structures, such as reference to nonsubjects. If so, people who read might be better at understanding unusual nonsubject references, or may be better at following cues that mismatch the subject bias (e.g., gaze cues). This possibility would be broadly consistent with findings that exposure facilitates the comprehension of rare syntactic structures (Fine & Jaeger, 2013; James et al., under review; Montag & MacDonald, 2015).

On the other hand, exposure may instead facilitate the development of representations about the difference between references in subject and nonsubject position. If readers have a more robust representation that subjects entities are likely to be mentioned, they may access this information more quickly or more reliably during language comprehension. That is, if a person has encountered a high volume of linguistic input, or input with greater complexity of utterances, they may develop a stronger representation of relevant patterns. This should lead to a more

systematic reliance on linguistic cues overall, and specifically with patterns like the high likelihood of reference to subject entities.

The Current Study

The goal of the current project is to test whether biases during pronoun comprehension are influenced by individual differences in language exposure. To test pronoun comprehension, we examine how adults interpret pronouns as a function of subjecthood and gaze cues, using Nappa and Arnold's video task in Exps. 1 and 2, and a variation of the task in Exp. 3. Nappa and Arnold (2014) asked adult participants to watch short videos in which a woman told stories about either two male or two female characters, where the characters' genders were learned at the start of the experiment (see Fig 1). On each trial, the speaker introduced the two characters, and then uttered a short, 2-sentence story as in (3).

(3) Panda Bear is having lunch with Puppy. He wants a pepperoni slice.

On critical trials, the question, e.g. "who wants the pepperoni slice?", revealed participants' interpretation of the pronoun. On the neutral trials, where the speaker gazed or pointed at the pizza, participants tended to choose the subject (the first-mentioned character) as the referent of the pronoun about 80% of the time. Yet when the speaker gazed or pointed at one of the characters, this bias shifted. Gazing or pointing at the subject increased the rate of choosing the subject as the referent of the pronoun to about 95%. When the speaker gazed at the nonsubject, responses dropped to chance, but when the speaker pointed at the nonsubject, listeners instead chose the nonsubject as the referent 90% of the time.

Figure 1. Example screen shot from Nappa and Arnold (2014) from the video corresponding to example 3, in the gaze-to-subject condition.

Nappa and Arnold's results highlight the fact that the effect of the linguistic context is variable. Moreover, there was substantial variation amongst individual subjects: The rate of choosing the subject character in the neutral conditions ranged from 2/8 to 8/8 (avg. 6.2). These stimuli provide little other information to guide pronoun interpretation, because the predicates were all equally applicable to the two characters. Thus, the neutral condition is a good estimate of individual variability in the usage of the subject bias.

The current study examines whether individuals vary in their reliance on the linguistic context to interpret pronouns, and whether this variation correlates with individual variability in linguistic experience. We focus only on the gaze conditions, which were less constraining than the pointing conditions, and thus more likely to reveal individual differences. This presents an opportunity to examine two questions. First, is sensitivity to the subject bias overall a function of language experience? We test each subject's preference to link pronouns with the grammatical subject character, and ask whether individual biases correlate with reading experience. Second, do individual differences lead to variable sensitivity to linguistic vs. social cues? That is, do some individuals show a preference for following gaze cues over linguistic cues and vice versa? We assess these questions primarily by analyzing participants' responses to the question "Who wants [the object]?", which signals their pronoun interpretation. As a secondary measure, we

analyze reaction time. Although we did not tell participants to respond quickly, reaction time is a broad measure of ease of responding. If print exposure facilitates use of the context, we might expect participants with high print exposure scores to respond more quickly.

Measuring individual differences in linguistic input presents a challenge to researchers. In particular, it is difficult to test variation in spoken language input, without access to an individual's history of daily interactions. We therefore approached this question by instead testing variation in exposure to written language. There are several reasons why this approach is advantageous. First, individuals are likely to have substantial variation in the quantity of written language exposure, independent of their use of spoken language. Children do not typically learn to read until they are school age, and adults vary in both reading skill and reading enjoyment.

Second, it is relatively easy to obtain a proxy measure of variation in print exposure, using the Author Recognition Task (Stanovich & West, 1989). In this task, participants are asked to indicate which authors they recognize from a list of fiction authors that includes both real authors (e.g., J.R.R. Tolkien), and fake author names. Performance on the task is measured as the number correct minus the number incorrect selections. This metric reflects exposure to written material, since people who read more tend to recognize more authors. The validity of this measure comes from evidence that the ART correlates with numerous measures of reading, including vocabulary knowledge, verbal comprehension, word identification, word naming, and reading speed (Mol & Bus, 2011; Moore & Gordon, 2014). The ART correlates with both self-reported measures of reading speed (Acheson, Wells, & MacDonald, 2008), and gaze duration in eyetracking studies (Gordon, Moore, Choi, Hoedemaker, & Lowder, under revision; Moore & Gordon, 2014).

In sum, we hypothesize that variation on the ART assessment of print exposure will predict performance on our offline pronoun comprehension task. Note that this approach is not designed to test whether written or spoken language is more important for the development of pronoun comprehension strategies, even though we have hypothesized that written language input may be especially relevant. Instead, any effect of language exposure will support the broader hypothesis that language input affects the mechanisms used for pronoun comprehension.

The hypothesized pattern is inherently correlational: participants with higher print exposure should have a stronger subject bias in pronoun comprehension. This raises questions about whether any observed patterns can be explained by other individual differences, and how both print exposure and pronoun comprehension relate to other demographic differences. We therefore additionally test working memory and theory of mind (Experiment 1), as well as reading skill and socioeconomic status (Experiments 2 and 3).

Experiment 1

We tested the relationship between print exposure and pronoun comprehension, using the stimuli from Nappa & Arnold's (2014) video task, with just the three gaze conditions. We also tested the effects of a) working memory and b) theory of mind.

We do not a priori expect strong effects of working memory, given that our stories and task are very simple, and working memory effects may be limited to longer texts (Daneman & Carpenter, 1980; van Rij, van Rijn, & Hendriks, 2011), or on-line measures (Nieuwland & van Berkum, 2006). However, if we do find any effects of print exposure, it is worth testing how they relate to working memory. Previous work has often tested working memory with sentence span measures, but this measure may also be influenced by variation in language exposure (see

discussion above). Therefore, to test individual differences in working memory we used a nonlinguistic memory task, the Automated Operation Span (Unsworth, Heitz, Schrock, & Engle, 2005).

As a second control measure we asked whether individual differences in print exposure were correlated with Theory of Mind, which is the ability to represent the knowledge, beliefs, and intentions of others (Premack & Woodruff, 1978; Wimmer & Perner, 1983). Researchers have suggested that Theory of Mind processing is related to both reading exposure (Kidd & Castano, 2013) and pronoun comprehension. Both van Rij et al. (2013) and Kehler and Rohde (2013; Kehler et al., 2008; Kehler, 2007) developed models of pronoun comprehension that involve representations of the speaker's preference for linguistic form, which could potentially (although not necessarily) index representations of the speaker's mental state. In addition, some authors have proposed that the production of pronouns is driven by the speaker's assessment of common ground knowledge (e.g., Chafe, 1976; Gundel et al., 1993). Alternatively, theory of mind scores may reflect attention to social cues such as gaze. We therefore tested theory of mind with Baron-Cohen et al.'s (2001) Reading the Mind in the Eyes task, which assesses an individual's ability to recognize emotions from images of eye expressions.

Methods

Participants.

A total of 72 native speakers of English participated at the University of North Carolina, Chapel Hill, in exchange for course credit. All participants in this and the other experiments in this paper provided informed consent. 11 participants were excluded for failing to meet accuracy criterion on the pronoun task (see below for description). One of these also failed to meet

criterion on the Automated Operation Span. This left 61 participants in the final analysis; respectively 24, 20 and 17 participants in lists 1, 2 and 3.

Procedure and Tasks.

Subjects participated in four tasks, total time approximately 50 minutes per subject. The order of the tasks was fixed: (1) Reference Task, (2) Automated Operation Span, (3) Reading the Mind in the Eyes and (4) Author Recognition Task. All the tasks were executed using the experimental software E-Prime 2.0.10. After completing the tasks, participants filled out a voluntary background questionnaire. Examples for this and all experiments in this paper are available at https://arnoldlab.web.unc.edu/publications/supporting-materials/supporting-material-for-arnold-strangmann-hwang-zerkle-nappa/.

Ambiguous pronoun interpretation task.

Each participant viewed a total of 54 videos (22 experimental stimuli, 32 fillers). In each video (see Fig. 2) participants saw a woman sitting at a table with two toy animal characters, one on either side of her, and a toy object in the center of the table. At the start of the task, the genders of the four puppets used were explicitly identified. In each story, the woman in the video introduced the two toy characters, and told a story about them (see example 4). On the next screen, a question appeared asking about which character wants the object in the center of the table, and the final screen asked participants to rate the naturalness of the videos on a scale of 1-7 (see Fig. 2).

Figure 2: Illustration of the Reference Task. Panel A depicts the gaze-to-subject condition; Panel B depicts the neutral gaze condition; Panel C depicts the gaze-to-nonsubject condition.

4) Example story for experimental stimuli: This story is about Puppy and Panda Bear. Puppy is having some pizza with Panda Bear. He wants the pepperoni slice.

QUESTION: Who wants a pepperoni slice?

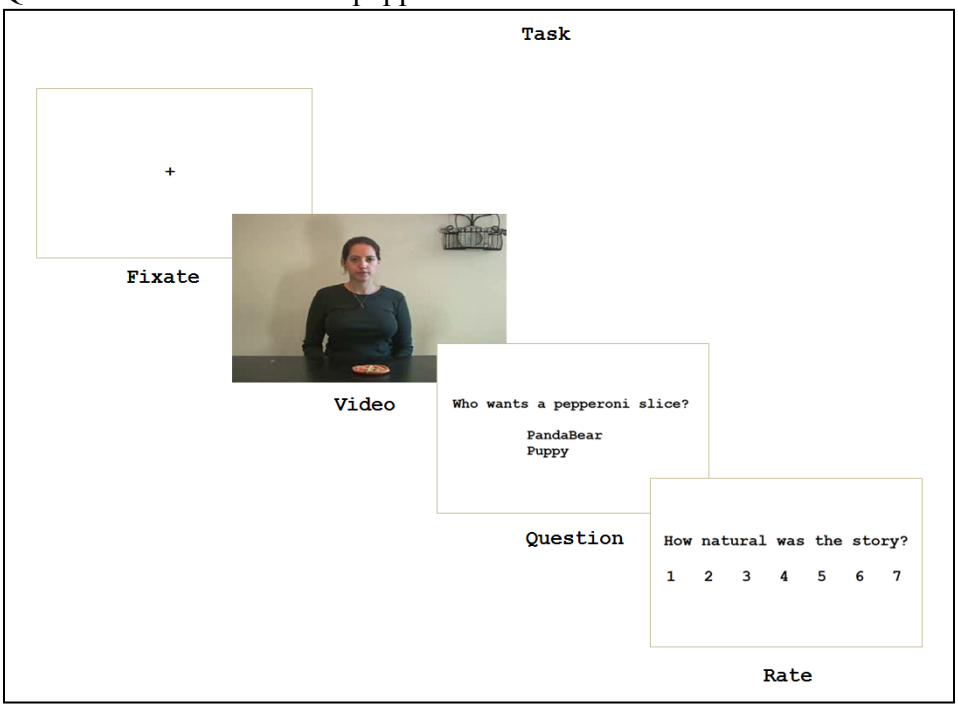

Figure 3. Sequence of the steps in the pronoun task: 1) fixation cross, 2) video, 3) story question, 4) naturalness rating.

Subjects viewed 22 experimental videos[1], viewing each story once in one of three conditions: 1) Gaze-to-subject (first-mentioned/subject character); 2) Neutral gaze cue to object; 3) Gaze-to-nonsubject (second-mentioned character). The gaze manipulation reflected the speaker's direction of gaze at the moment of uttering the pronoun. This was not a subtle cue, and involved changes to both eye direction and head/body direction.

[1] We only used 22 out of the original 24 experimental items due to technical problems with two of the items.

This design allowed us to assess how listeners differ in their reliance on cue type. The primary cue available to listeners in the neutral condition was the first-mentioned/subject bias. There were no other discourse status differences between characters, and we did not expect our stimuli to give rise to substantial real-world biases. The gaze conditions added a social cue that sometimes coincided with the linguistic context (looking at the subject), and sometimes conflicted with the linguistic context (looking at the nonsubject). For the experimental stimuli, the following screen asked which character wanted the mentioned object (e.g. the pepperoni slice) and the participant responded by pressing a keyboard key. We analyzed the proportion of the subject responses, i.e. those that support the discourse-based subject bias.

There were also 32 filler videos. These followed a similar format as the experimental stimuli, except the two characters were sometimes of different genders, and the second sentence usually mentioned one of them by name, instead of using a pronoun. In addition, all of the fillers included both gazing and pointing to the character who was being referred to. 16 of the filler items questioned the object (4), and 16 questioned the location (5):

4) Object Filler Example: This story is about Froggy and Bunny. This is Froggy and this is Bunny. Froggy is watching a movie with Bunny. Froggy wants the popcorn.
QUESTION: What does Froggy like to eat?

5) Location Filler Example: This story is about Panda Bear and Puppy. This is Puppy and this is Panda Bear. Puppy is playing in the sandbox with Panda Bear. Puppy wants the shovel.
QUESTION: Who was on the left hand side of the screen?

The object fillers offered two answer options: the mentioned object (e.g. popcorn) and a foil (e.g. McNugget). The location fillers had two possible answers (e.g., Puppy, Panda Bear), and the correct answer was counterbalanced. Two participants asked whether the location question should be answered from their own perspective, and they were both told that they should.

The fillers fulfilled several functions. First, they increased the variation in the story types, reducing the proportion of trials with ambiguous pronouns. Second, the location fillers required participants to watch the video, which ensured that they would see the gaze cues. Third, the fillers were used to assess whether participants were paying attention to the stories and images. Participants who answered incorrectly on more than 4 of the fillers were excluded from the analysis.

In the instructions, subjects were told that they would participate in a task designed for preschool-aged children, in order to explain the child-like nature of the stimuli, and encourage them to approach the task in a straightforward way. They were briefly introduced to the 4 characters: Puppy and Panda Bear (identified as male) and Froggy and Bunny (identified as female). We told participants to pay close attention to the videos, since we would be asking questions about them. Following the questions, subjects were also asked to rate the story on plausibility (1 being not plausible at all and 7 being very plausible). These ratings were included for the sole purpose of drawing the participant's attention away from the actual purpose of the experiment, and were not analyzed.

Working memory: Automated Operation Span.

To measure working memory, we adopted a task that was developed by Unsworth, Heitz, Schrock, and Engle (2005), which is distributed as a pre-programmed E-Prime task. The main task presented participants with a math problem. On the next screen participants had to indicate whether the answer was correct or not. Lastly, they saw a letter, which they had to remember. This math problem/letter sequence was repeated 3-7 times, resulting in 3-7 letters to be

remembered. A recall screen asked participants to recall all the letters, with their score reflecting the number of letters recalled in the correct serial order.

Prior to the main task, participants were given three practice sessions. 1) Letters. In the first practice session, participants saw individual letters appearing on the screen; then a recall screen appeared, showing 12 different letters in 4 x 3 matrix. Participants were instructed to recall the letters in the same order they were presented by clicking the box next to the appropriate letters. The letter recall practice session included 8 trials. 2) Math. In the second practice session, fifteen simple math problems appeared one by one on the screen. On the next screen a digit was presented, and participants indicated whether this digit was the answer to the previous math problem or not, by clicking "true" or "false". The math practice session had two goals: 1) to acquaint subjects with the actual test and 2) to calculate the average time each individual subject needed to solve math the operations. This individualized time was subsequently used as a personalized time limit in the actual test to solve the math problems, in order to minimize the possibility of letter rehearsal during the math task. If a subject did not manage to solve a math problem within their personalized time limit, the task automatically continued to the letter screen, counting that item as an error. 3) Math and Letters. The third practice session was similar to the actual task. Participants performed both math problem solving and letter recall. The third practice session included three practice trials, each of set size 2.

In the actual test, set sizes ranged from 3 to 7 math problems/letters. The order of set sizes was randomized for each participant. Participants were instructed to keep their overall math score at 85% or above, which they could see during the feedback. This 85% norm (64 correct out of 75) was adopted from Unsworth et al. (2005) and was used as a criterion to exclude

participants from analysis, since we were only interested in participants who actually attempted to solve the math operations.

The Automated Operation Span yields two scores: 1) the partial storage score, which reflects the sum of all trials in which the letters were recalled in the correct serial position, and 2) the absolute score, which counts only those trials on which all letters were recalled correctly (Redick et al., 2012). We used the partial storage score because it is more a sensitive measurement of working memory capacity compared to the absolute score (Redick et al., 2012).

Theory of mind: Reading the mind in the eyes revised.

To assess Theory of Mind, we used a digitized version of the Reading the Mind in the Eyes task (Baron-Cohen, Wheelwright, Hill, Raste, & Plumb, 2001). Each participant viewed a total of 36 pictures of eyes and one practice picture, all expressing human emotions. The participant's job was to indicate which one of four words (emotions) they felt best described what the person in the picture was either thinking or feeling, by pressing one of four marked keys. One of the four words was the target answer, the other three were foils. If a subject did not know what a word meant, they could look it up in the word definitions list (this happened rarely). Participant responses were analyzed for accuracy based on the key in Baron-Cohen et al. (2001).

Print exposure: Author Recognition Task.

We assessed print exposure using a digitized version of Moore and Gordon's (2014) version of the Author Recognition Task (ART). This version uses the 65 author names from Acheson, Wells and MacDonald (2008), and the 65 non-author foils from the Martin-Chang and Gould (2008) adaptation of Stanovich and West's original ART (1989).

Subjects saw 130 names on the screen, 65 of which were authors. Participants were asked to mark the names they recognized as authors. Each subject's score was the total number of correct authors marked, minus the number of foils marked. Our E-Prime version of the task used the same name order as in Moore and Gordon's study, and participants selected each name by clicking a checkbox right next to each individual name. To prevent participants from random guessing, we informed them that their score would be penalized for checking non-authors. Subjects were not timed, but were told to try to do the task as fast as possible.

Results and Discussion

Reference Task Analysis.

We first examined participants' responses in the pronoun task, which asked them which character wanted the object. Participants were most likely to select the subject character. However, this general bias was modulated by the gaze manipulation: the subject was chosen on 93% of trials when the speaker gazed at the subject, 86% of trials on neutral-gaze trials, and 67% of gaze-to-nonsubject trials. This closely matched the pattern for these conditions in Nappa & Arnold (2014, Experiment 1; 93%, 79%, and 50%, respectively).

We assessed the statistical significance of this pattern with a mixed effects logistic regression, using SAS proc glimmix with a binary distribution and a logit link. Our dependent measure was the choice of subject vs. the nonsubject. The gaze predictor was modeled as a 3-way categorical variable, parameterized as two dummy variables (gaze-to-subject: 1 vs. 0 and gaze-to-nonsubject: 1 vs. 0). All models included crossed random effects for both participant and item and random slopes for critical predictors with respect to both participant and item (as appropriate). When the model estimated a random effect to be zero, it was removed from the model. For each model, we first built a control model to assess the contribution of design

variables (left/right location of the subject in the stimulus video; top/bottom location of the subject name on the question page; List 1 vs. 2 vs. 3 (parameterized as two binary predictors); male vs. female stimulus characters, and trial order). Any control variables that had a t-value greater than 1.5 in the control model were retained for the final model. In the final responses model, only the dummy variables for List were included.

The model revealed significant effects of both gaze-to-subject ($\beta = 0.98$ (SE = 0.33), $t=3.0$, $p = .004$) and gaze-to-nonsubject ($\beta=-1.27$ (SE = 0.23), $p < .0001$)[2]. The significance of each variable indicates that the rate of choosing the subject was higher for the gaze-to-subject condition than the neutral condition and that the rate of choosing the subject was lower for the gaze-to-nonsubject than the neutral condition. The control List variables were also significant.

Individual differences.

We first identified the mean, standard deviation, and range of responses for our individual differences tasks. Table 1 demonstrates that our data are very similar to data acquired by the original task developers.

We also examined the correlations amongst these individual predictors, and found that none were significantly correlated with each other: The Author Recognition Task vs. the Automated Operation Span ($r = .006$, $p = .96$); Author Recognition Task vs. Reading the Mind in the Eyes ($r = .04$, $p = .75$), and Operation Span vs. Mind in the Eyes ($r = .11$, $p = .39$).

[2] This model contained random intercepts for both subject and item, and random slopes for gaze-to-subject by subjects, and gaze-to-nonsubject by both subjects and items. The slope for gaze-to-subject by subjects was estimated to be zero by the model.

Table 1: Descriptive statistics showing mean and *SD* for our study and data from the original task developers.

Measurement	Task	Mean	*SD*
Working memory Automated Operation Span	Our study	56.39	14.12
	Redick et al. (2012)[i] N = 6,236	57.36	13.65
Theory of Mind Reading the Mind in the Eyes	Our study	28.98	3.20
	Baron-Cohen et al. (2001) [ii] N = 103	28.0	3.5
Print exposure Author Recognition Task	Our study	15.43	7.46
	Moore & Gordon (2015) [iii] N = 1012	14.72	7.32

i. The data shown from Redick et al.'s (2012) study are the partial scores of the overall population.
ii. Data from Baron-Cohen et al.'s (2001) study represent their student population, which most closely resembles our study's population.
iii. Data shown from the Author Recognition Task by Moore & Gordon (2015) are data derived from the 65 author scale with a standard ART score.

Our focal predictor, the Author Recognition Task (ART), is known to correlate with other standardized measures of language processing, like the SAT (Standardized Aptitude Test). 40 of our participants reported their SAT verbal score from memory on a voluntary background questionnaire, and this score correlated with the ART at $r = .48$, $p = .0016$. This confirms the validity of our ART measure as an indicator of individual differences in verbal experience and skill, even though the voluntary report of SAT scores may be only approximate for some participants, since their memory may not have been accurate.

The critical question in the current study was whether individual differences would affect pronoun interpretation or modulate the effect of gaze. We examined this question by separately adding each of our metrics (print exposure, working memory, Theory of Mind) to the response model described above, as a centered predictor. In each case, we then built an additional model to test for interactions between each individual difference metric and the gaze manipulations.

However, in no case were there any significant interactions, so here we report only the models with main effects. These analyses revealed that there were no effects of either our working memory metric or our Theory of Mind metric. However, we did find strong effects of participants' print exposure, as measured by the Author Recognition Task.

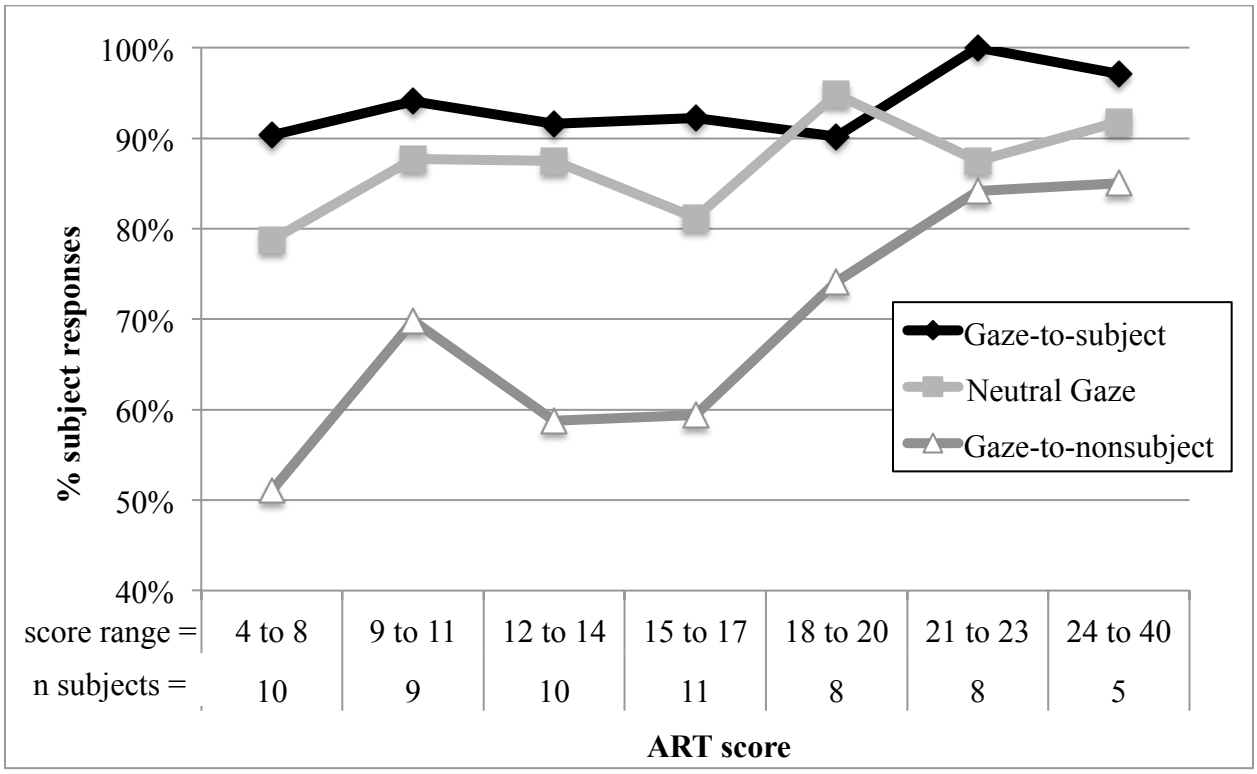

Figure 4: Individual differences in the average subject response for each gaze condition, depending on individual score on the ART task. Participant means are averaged in bins according to ART score, ranging from the lowest bin (score 4-8) to the highest (score 24-40). Each bin represents 3 ART values except the highest and lowest bins.

Figure 4 plots the effects of ART on participants' responses, by binning participants by their ART scores (on the x-axis). This figure illustrates that participants with high ART scores tended to choose the subject significantly more often than participants with lower ART scores. This effect appears strongest for the gaze-to-nonsubject condition. However, in our overall model, there was no significant interaction between either gaze predictor and ART. In other words, participants with higher ART scores chose the subject more systematically overall.

Table 2 reports the results from the analysis including print exposure. This analysis again revealed a main effect of our gaze manipulation, as well as a significant effect of print exposure, such that participants with greater print exposure had a higher rate of the subject responses. Additional models on the response data confirmed that ART did not interact with the gaze manipulation. In separate models, we examined the effects of theory of mind and working memory. In each model, the gaze predictors had similar values, but the individual differences predictors were not significant (Reading the Mind in the Eyes: β= 0.03 (SE= 0.05), t= 0.64, p = 0.53; Automated Operation Span: β = 0.005 (SE = 0.01), t = 0.39, p = 0.70). [3]

Table 2. Responses analysis with print exposure: Parameter estimates and statistics from the analysis of the subject responses, including print exposure as a predictor.

Predictor	Estimate (SE)	t	p
Gaze-to-subject	0.98 (0.33)	2.95	0.004
Gaze-to-nonsubject	-1.27 (0.23)	-5.62	<.0001
Print exposure (ART)	0.06 (0.02)	2.3	0.03
List 1	0.93 (0.41)	2.28	0.03
List 2	1.47 (0.43)	3.46	0.001

The pattern in Figure 4 seems to suggest that print exposure had the largest effect on the gaze-to-nonsubject condition, where the gaze cue conflicted with the linguistic context. Yet the interaction term was not significant. Nevertheless, it appears that the main effect of print exposure was carried by the gaze-to-nonsubject condition. This trend was confirmed by the fact that print exposure correlated with the average subject response for the Gaze-to-nonsubject condition (r = .29, p = .02), but not the other conditions (r's < .19, p's > .16). This suggests that

[3] All of the response models including individual difference predictors had a random intercept for both subject and item, and a random slope for gaze-to-nonsubject by both subject and item, and a random slope for gaze-to-subject by subject. The slope for gaze-to-subject by subject was estimated to be zero by the model. The random intercept for item was also estimated to be zero, but was retained because the slope was nonzero. The ART model estimated the ART x item slope to be zero so it was removed.

the higher ART participants may have experienced less indecisiveness about responding, especially in the condition with the greatest cue conflict.

Consistent with this interpretation, we found that participants with high print exposure responded more quickly. We analyzed the time to respond from the onset of the question screen, excluding outliers greater than 3 standard deviations from the mean (1.8% of the data). We used a mixed effects linear regression, using SAS proc mixed, with log of the reaction time as the dependent measure. We first tested control predictors in a separate model, retaining only those that were significant at $|t| > 1.5$. Only item order was retained for the final model. We added our two binary (centered) gaze predictors, whether the response selected the subject character or not, item order, and ART score (centered). As shown in Table 3, latencies were faster for the subject responses, later items in the task, and participants with higher ART scores. Additional models confirmed that ART did not interact with condition or response. Thus, high print exposure led to fast responses overall.

Table 3. Experiment 1 Reaction time analysis[4]

Predictor	Estimate (SE)	t	p
Gaze-to-subject	-0.01 (0.01)	-1.1	0.27
Gaze-to-nonsubject	0.02 (0.01)	1.7	0.09
Subject response	-0.03 (0.01)	-2.11	0.04
Item order	-0.003 (0.0003)	-13.88	<.0001
Print exposure (ART)	-0.005 (0.002)	-2.35	0.02

In sum, the results of Experiment 1 demonstrate that individual differences in print exposure predict pronoun comprehension, such that people with greater reading exposure tended to select subject referents more often during spoken pronoun comprehension. We also found that this effect was not explained by either working memory or theory of mind differences. While

[4] The reaction time model had a random intercept for both subject and item, and random slopes for gaze-to-subject and gaze-to-nonsubject by subject, and for ART by item; the slopes for both gaze-to-subject and gaze-to-nonsubject by item were estimated to be zero by the model.

other studies have reported correlations between the ART and the Reading the Eyes in the Mind task (Panero et al., 2016; Samur et al., in press), in our sample there was no hint of a relationship ($r = .04$).

Our results suggest that exposure shapes discourse processing, much like it affects syntactic and lexical processing. Although exposure facilitates the processing of rare words and syntactic structures, we found that for pronoun comprehension it supports the use of the most frequent pattern. We return to this issue in the general discussion.

We observed an effect of experience here by measuring print exposure, as indexed by the ART. The ART score does not directly measure amount of time spent reading, in that people also gain exposure to authors and printed materials by hearing books read aloud, or talking about books (Stanovich & West, 1989). However, reading is certainly the most obvious way of gaining print exposure, and it is likely that much of our individual variation stems from differences in time spent reading. This raises questions about how the ART relates to other individual differences that are likely to correlate with reading, namely differences in language skill and socioeconomic status. We did not report the effects of SAT scores in Experiment 1, because we were not confident in subjects' SAT reports by memory. We test these variables more explicitly in Experiment 2.

Experiment 2

We repeated the task in Experiment 1, with a few changes. First, we measured verbal skill and SES predictors instead of theory of mind and working memory, which were nonsignificant in Experiment 1. Second, we used a task with fewer items (15 critical, 10 fillers), which revealed a similar pattern of responses in a study with children (Arnold, Castro, Zerkle, &

Rao, in preparation). Third, we used a slight variant on the ART task. Fourth, we omitted the plausibility rating.

Methods

Participants.

A total of 71 native speakers of English participated at the University of North Carolina, Chapel Hill, in exchange for course credit. 12 participants were excluded because they had previously participated in another study that also tested the ART. Two participants were excluded because E-prime froze during the pronoun task. One participant was excluded because they reported a reading disorder on the background questionnaire. This left 56 participants in the analysis, 20 on list 1 and 18 each on lists 2 and 3.

Procedure and tasks.

Subjects began with the pronoun task, which was a shortened version of the task used in Experiment 1. They then filled out a questionnaire in Qualtrics, which asked about their background information, and included four measures critical to the study: a) their SAT scores, b) the Shipley vocabulary test, c) questions about socioeconomic status, and d) the ART. Total time of participation was approximately 30 minutes per subject.

Pronoun task.

The pronoun task was the same as the one in Experiment 1, except we included fewer trials, and there was no plausibility question. There were 15 experimental items (5 in each of the three gaze conditions), and 10 fillers (6 location fillers and 4 object fillers).[5] Given the low number of fillers, we did not use location fillers (which are harder) as a diagnostic for exclusion. Nevertheless, performance was high overall (98% correct), and no participant missed more than 4 location fillers; performance on the object fillers was 100% correct.

[5] This shorter version of the task is identical to the one we used with children (Arnold et al., in preparation).

SAT scores.

Participants were asked to indicate if they had taken the SAT. The University of North Carolina requires applicants to report scores for either the SAT or the ACT, so not all students have taken the SAT. If they did, we asked them to look up their scores, so they would be accurate. 50 of the 56 subjects in our analysis reported SAT scores; 38 of these looked up the scores and 12 reported them from memory if they could not access their scores.

Shipley Institute of Living Scale: Vocabulary.

Participants' knowledge of vocabulary was assessed by asking them to identify the synonym of 40 words (Shipley, 1940). In each case, they were presented with 5 options. The words became harder and harder as the test progressed, in part because the correct answer was sometimes a synonym of a subordinate meaning of the word.

Socioeconomic status.

We asked three questions[6] about socioeconomic status: 1) Family Income: participants selected the range representing their family's income, to the best of their knowledge. 21 range options were provided (less than 50,0000; 50,001-100,000; 100,091-150,000….. over 1,000,000). 2) Mother's education, and 3) Father's education. For both parental education questions, they selected one of the following options: did not graduate high school; graduated high school; some college or 2-year degree; 4 year degree; Graduate or professional degree; doctorate; do not know).[7]

[6] We also asked subjects to subjectively rank their position on a socioeconomic ladder (B. K. Payne, unpublished data, U. Chapel Hill). However, due to a formatting error we omitted this question from analysis.
[7] For analysis, we combined "less than high school" with "high school", and "graduate/professional" with "doctorate"

Author Recognition Task.

We used a task that was nearly identical to the ART used in Experiment 1. This version had only 62 authors and 64 nonauthors, and used a slightly different selection of names. This variation of the task was developed by Peter Gordon's lab (personal communication), for the purpose of designing a task with greater sensitivity for the UNC undergraduate population. They replaced some author names that were almost never selected (Moore & Gordon, 2015) with better-known authors, and replaced some nonauthor names with less confusable names.

Self-report of language exposure activities.

We gathered additional information about individual participants' personal activities, by asking them to rate the following questions:

- Not including school assignments, in a typical week, how many hours (on average) do you read books?
- Not including school assignments, in a typical week, how many hours (on average) do you browse internet sites?
- Not including school assignments, in a typical week, how many hours (on average) do you listen to books read aloud?
- How much do you enjoy reading?

Results and Discussion

Reference Task Analysis.

As in Experiment 1, we found a general preference to select the subject, as well as gaze effects. Participants selected the subject 94% in the neutral gaze condition, 96% in the subject-gaze condition, and 66% in the nonsubject-gaze condition. We again modeled this with a logistic

regression, with random intercepts for both subjects and items, and random slopes for both gaze-to-subject and gaze-to-nonsubject predictors by subjects. We found a significant effect of gaze-to-nonsubject (β = -2.35 (0.35), t = -6.8, p <.0001). There was no effect of gaze-to-subject (β = 0.38 (0.47), t = 0.81, p = 0.42).[8]

Overall, participants chose the subject much more often than in Experiment 1. This may have been a result of the fewer number of filler items. All of the fillers involved pointing, even though pointing was not necessary to resolve an ambiguity, as all fillers used names. We speculate that the presence of pointing in Experiment 1 may have drawn attention toward the social cues, and away from the linguistic context. With fewer pointing fillers in Experiment 2, participants may have been more likely to focus on the linguistic context.

Individual Differences.

Table 4 reports the average and range for each of our individual difference variables. We first tested each predictor separately, by adding it to the main model described above. The critical question was whether ART scores would predict responses. For the SES and skill measures, we first tested each predictor in a separate model. Any predictor that was significant or marginal was then added to the model with ART, to directly test the relative importance of each predictor.

Our first question was whether print exposure would predict individual variation in pronoun comprehension. Figure 5 shows that we observe the same general trend as in Experiment 1, with higher subject responses for participants with higher ART scores. When we added ART to the model, we found the same gaze effects (gaze-to-subject (t = 0.86, p = 0.39);

[8] In this model, the slopes for gaze-to-subject and gaze-to-nonsubject by item were predicted to be zero by the model and removed.

gaze-to-nonsubject (t = -6.7, p < .0001) and in addition we found a nearly-significant effect of ART score (β = 0.061 (0.03), t = 1.98, p = 0.053).[9]

Table 4. Individual difference predictors in Experiment 2

		average	range
PRINT EXPOSURE	ART score	16.8	3-32
LANGUAGE SKILL	Shipley vocabulary test	31.3	21-37
	SAT reading score	633.3	460-780
SOCIOECONOMIC STATUS	mother's education	2.1	1-4[10]
	father's education	2.2	1-4
	Income	16.5[11]	1-21

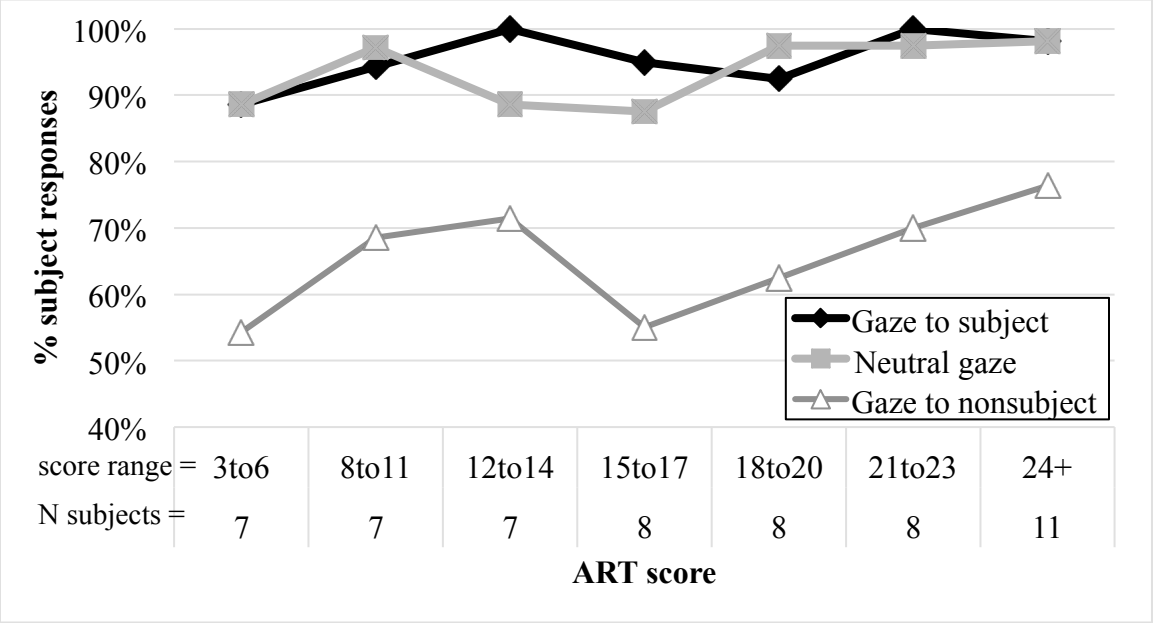

Figure 5. Percentage of subject responses in Experiment 2, binned by ART score.

[9] The random slopes for gaze-to-subject and gaze-to-nonsubject by items were estimated to be zero and removed. The ART slope by items prevented the model from converging and was excluded.
[10] 1 = HS or less; 2 = some college; 3 = 4 year college; 4 = graduate school/professional school/doctorate.
[11] This annual income rank falls between the bins 201K-250K and 251K-300K. Responses ranged from less than 50,000 to over 1,000,000.

We again examined reaction times in a mixed effects linear regression model, following the same modeling procedure as in Experiment 1, and again excluding responses that were longer than 3 standard deviations from the mean. Average RT in this experiment was somewhat slower than in Experiment 1 (*M* = 3013 ms vs. *M* = 2096 in Experiment 1). A control model revealed that item order was a significant predictor (later items were faster), and responses were faster when the subject option was on top; these two predictors were retained for the final model. We built a model with two binary (Gaze-to-subject and Gaze-to-nonusbject) centered predictors for condition, a binary predictor for the response (subject response vs. nonsubject response), item order, ART score (centered), and the interaction between ART score and subject response. We initially tested three interactions with ART (ART x gaze-to-subject, ART x gaze-to-nonsubject, and ART x subject response), but only retained interactions that were significant at $|t| > 1.5$ for the final model (here, ART x subject response).

As shown in Table 5, the final model showed that people were slower in the gaze-to-nonsubject condition (*M = 3296)*, compared to the other two conditions (gaze-to-subject *M* = 2981; neutral gaze *M* = 2855). They were also slower when they selected the nonsubject as their response, and for earlier items. We also replicated the finding that participants with higher ART scores responded more quickly.

Table 5. Experiment 2 reaction time analysis[12]

Predictor	Estimate (SE)	t	P
Gaze-to-subject	0.001 (0.012)	0.05	0.962
Gaze-to-nonsubject	0.059 (0.013)	4.61	<.0001
Subject response	-0.019 (0.023)	-0.82	0.419
Item order	-0.008 (0.001)	-8.44	<.0001
Subject option on top	-0.038 (0.018)	-2.14	0.052
Print exposure (ART)	-0.008 (0.003)	-2.6	0.013
ART * subject response	0.004 (0.003)	1.33	0.191

[12] The random slopes for gaze-to-subject and Subject option of top by participants, and for ART and gaze-to-nonsubject by items were estimated to be zero by the model and removed.

Our second question concerned measures of language skill, and how they would relate to both ART and pronoun comprehension. ART correlated with both reading SAT ($r = 0.45$, $p = .001$) and Shipley vocabulary score ($r = 0.48$, $p = .0002$). We tested vocabulary and reading SAT (for the 50 participants who reported it) as predictors in separate models, in each case adding these to the basic model with the gaze predictors. Reading SAT had no effect ($\beta = .002$ (.002), $t = 0.72$, $p = 0.47$)[13], and the Shipley vocabulary predictor had a marginally significant effect ($\beta = 0.13$ (0.07), $t = 1.82$, $p = 0.075$)[14]. If we add Shipley scores to the model with ART, neither Shipley nor SAT have significant effects (t's < 1.5, p's $> .2$). Thus, the ART effect may be related to language skill, but vocabulary knowledge does not overshadow ART as a stronger predictor.

Finally, we asked whether pronoun comprehension was predicted by socioeconomic status, which could correlate with both amount of time spent reading and other educational opportunities. However, none of our three measures of SES were correlated with ART (all $|r|$'s $< .1$; all p's > 0.6). We added each one separately to our basic model with the two gaze predictors. None of the SES variables had significant effects, although mother's education was marginally significant ($\beta = -0.25$ (0.15), $t = -1.67$, $p = 0.010$)[15]. However, note that the effect is the opposite of what might be predicted – participants with more educated mothers had a slightly lower likelihood of selecting the subject. If we add mother's education to the model with ART, we find

[13] The random slopes for gaze-to-subject, gaze-to-nonsubject, and SAT by items were estimated to be zero by the model and removed.
[14] The gaze-to-subject, gaze-to-nonsubject and Shipley slopes by items were estimated to be zero by the model and removed.
[15] This model had slopes for gaze-to-subject and gaze-to-nonsubjects by subjects, but both these and the Mother's education score slopes by items were estimated to be zero by the model.

similar effects as before,[16] but mother's education is not significant (β = -0.2 (0.15), t = -1.6, p = .11). Thus, print exposure effects are not explained by socioeconomic status.

To better understand our ART measure, we examined correlations between ART scores and responses to our questions about reading and internet use. ART score was correlated with the number of hours spent reading books (r = 0.37, p = 0.005) and how much people enjoy reading (r = -0.61, p < .0001)[17]. It was unrelated to internet browsing (r = 0.01). Although listening to books could potentially increase print exposure, 85% of participants reported that they listened to books "rarely or never", and this measure was also unrelated to ART scores (r = 0.02). These correlations support our interpretation of the ART scores as a proxy for actual reading behavior.

The results from Experiment 2 replicate the basic ART effect found in Experiment 1, although it was somewhat weaker here. The weaker effect of print exposure may be due to the fact that there was less variability in responses on this shorter task, compared to the longer task in Experiment 1. In addition, these findings demonstrate that comprehension may be related to other measures of comprehension skill, given that vocabulary was a marginal predictor of the subject bias, and adding it to the model with ART eliminated the significance of ART scores. In addition, both reading SAT scores and vocabulary scores were correlated with ART scores.

What type of mechanism would explain the exposure effects? Even though ART scores correlate with language skill, our pronoun comprehension findings cannot be interpreted as a simple effect of skill. First, linking the pronoun with the nonsubject is not a case of comprehension failure. Pronouns can grammatically and naturally refer to either the subject or the nonsubject. Even though the subject bias is strong, it is not categorical or grammatically

[16] Gaze-to-subject had no significant effect (t < 1.0; p > .3); gaze-to-nonsubject reduced subject choices (β = -2.4 (.35), t = -6.78, p < .001) and ART marginally increased subject choices (β = .06 (.03), t = 1.91, p = .061).

[17] This is negative because our responses were 1 = love it; 2 = like it; 3 = neither like nor dislike it, and 4 = not very enthusiastic about it, so this shows that higher ART scores correlate with more enjoyment of reading.

required. Second, our stories were linguistically simple, and designed to be understandable to preschool children. The vocabulary was therefore not difficult, and it is unlikely that vocabulary comprehension caused individual differences in pronoun comprehension. Thus, if vocabulary knowledge is not mediating the ART effect, it leaves open the possibility that print exposure also has independent effects on discourse processing biases.

We consider the hypothesis that language exposure, here measured by print exposure, is important for developing the knowledge that entities in subject position are likely pronoun referents. That is, this is not a uniform bias across all individuals. However, the pronoun task used in Experiments 1 and 2 raises an alternate possibility. In both cases, people with a greater subject bias could also be described as people who were less sensitive to the gaze cue. We did not find support for this hypothesis, in that ART score did not interact with the gaze condition. Nevertheless, the numerical trend suggests that people with high ART scores also showed less variation between the gaze-to-subject and gaze-to-nonsubject conditions. Does print exposure instead reflect sensitivity to the gaze cue? We tested this alternate possibility in Experiment 3.

Experiment 3

The results of Experiments 1 and 2 do not rule out the possibility that our results stem from variation in how people use gaze cues, and not from variation the strength of the subject bias. That is, it might be that people with high print exposure tend to ignore gaze cues, either because they deem them unimportant, or perhaps do not pay attention to them. In our video task, this leaves only the subject bias as a reasonable guide to pronoun interpretation. If so, we would expect the effect to disappear in an experiment without gaze cues. Alternatively, if print exposure

affects use of the subject bias itself, it should persist in a new task. We therefore designed a simpler version of the experiment, without any gaze or pointing information.

Methods

Participants.

A total of 66 native speakers of English participated at the University of North Carolina, Chapel Hill, in exchange for course credit. 4 participants were excluded; 1 was a non-native speaker of English, 2 reported language/hearing disorders, and we had technical difficulties for one. This left 62 participants in the final analysis.

Procedure and tasks.

Subjects began with the pronoun task, which was presented with the software Paradigm. They then filled out a questionnaire in Qualtrics, which asked about their background information, and also included two measures critical to the study: a) their SAT scores, and b) questions about socioeconomic status. They then were directed to a separate Qualtrics questionnaire that contained the Author Recognition Task. Total time of participation was approximately 30 minutes per subject.

Pronoun task.

Similarly to the video task in Experiments 1 and 2, participants heard simple stories about two characters, but without any video of the speaker. The stories were about four characters (Ana, Liz, Will, and Matt). We designed these stories with adult listeners in mind, so characters took part in activities like cleaning, eating at a restaurant, or going to yoga. There were 24 experimental stimuli and 36 fillers, all of a similar format to the stories in Experiments 1 and 2. The experimental stories were all about same-gender people, with each of the four characters in subject position in 6 stories. All sentences began with a sentence with the structure “X is doing

something with Y", with the first-mentioned character on the left half of the time. The second sentence started with a pronoun, and half used the verb "needs", half "wants". E.g., *Ana is cleaning up with Liz. She needs the broom.* The filler stories used a similar format, but continued with name references instead of pronouns. Half the fillers introduced the characters in a coordinate NP (e.g., *Will and Matt are getting ready for school. Matt needs a jacket*,) and half in the same with-structure as used in the critical stimuli (*Liz is going to a library with Ana. Ana wants to borrow a comic book*.)

All stories were followed by two questions, with a two-alternative forced choice answer. The critical question for experimental stimuli asked who needs/wants the object, and thus indicated pronoun comprehension. This critical question was sometimes the first question, sometimes the second, and the subject choice appeared on the top for half the trials. Half of the time the other question asked "What are they doing?", and half the time it asked "Who was on the left (or right) side of the screen?"

Filler items also occurred with three question types. The "Who needs the [object]" question" occurred 20 times, and served as a check to make sure that participants were paying attention. All fillers used names for the second sentence, so there was a single right answer. No participant missed more than one of these questions, with an average of 99% correct. The "What are they doing" question occurred 22 times, and no participant missed more than two of these questions, with an average of 99% correct. The location question "Who was on the left/right side of the screen" was harder, and thus was not used as a criterion for inclusion. Nevertheless, all subjects answered at least 15 correct, with an average of 21 correct (95%). For all the critical and filler questions, the two possible answers were shown vertically, and the participant pressed a

button to signal the top or the bottom answer. The correct answers occurred equally on top and bottom.

Language skill measures.

- **SAT scores.** These were elicited in the same manner as for Experiment 2. The Shipley vocabulary test was not included in this experiment.[18]
- **Socioeconomic status.** We used the same measures as in Experiment 2: mother's education, father's education, and estimate of family income.
- **Author recognition task.** The same task as in Experiment 2 was used.
- **Self-report of language exposure activities.** We asked the same questions as in Experiment 2.

Results

The average rate of selecting the subject across all participants was 94%, with a minimum of 62.5% and a max of 100%. A total of 26 participants selected the subject character 100% of the time. We additionally tested whether the ART, SAT reading, and SES variables predicted responses, using the same modeling procedure as for Exp. 2; see Table 6 for average individual responses. ART correlated with SAT reading scores[19] ($r = .69$, $p < .001$). It also correlated with mother's education[20] ($r = .30$, $p = .02$) and marginally with father's education ($r = .23$, $p = .08$), but no other SES variables (r's $< .05$).

[18] We began collecting data for Experiment 3 before Experiment 2.
[19] Of our 62 participants, 53 reported SAT scores. 2 of these were in the wrong or inconsistent format and excluded from the SAT analysis. One reflected the new scale, and was converted to the old scale using conversion charts from df" https://collegereadiness.collegeboard.org/pdf/higher-ed-brief-sat-concordance.pdf. Of these 51 included participants, 6 reported their scores from memory.
[20] One participant reported mother's education as unknown and was excluded from analyses with this variable.

Critically, there was a significant effect of print exposure (β = 0.05 (0.02), t = 2.56, p = 0.014). We also found a significant effect of the SAT reading score (β =.006 (.002), t = 2.63, p = .01). When we included SAT and ART in the same model, neither had a significant effect (t's < .16, p's > .12).

As in Experiment 2, the family income and father's education predictors had no effect on pronoun interpretation (t's < 1.1, p's > .29). Mother's education was marginally predictive of pronoun response (β = .041 (.22), t = 1.83, p = 0.075)[21]. This time, unlike in Experiment 2, participants whose mothers had more education were likely to have somewhat higher ART scores. We added mother's education to the model with ART, which demonstrated that ART was a significant predictor of choosing the subject (β = .04 (.02), t = 2.09, p = .04), but mother's education was not (β = .27 (.23), t = 1.18, p = .24). This shows that like in Experiment 2, SES variables do not account for the ART effect.

Table 6. Individual difference predictors for Experiment 3.

		average	Range
PRINT EXPOSURE	ART score	20.31	1-44
LANGUAGE SKILL	SAT reading score	655	470-800
SOCIOECONOMIC STATUS	mother's education	3.03	1-4
	father's education	2.9	1-4
	income	17.77[22]	5-21

[21] One participant did not report mother's education and was excluded from this analysis and other analyses with this variable.

[22] This annual income rank falls between the bins 150K-200K and 200K-250K. Responses ranged from less than 50,000 to 800K-850K.

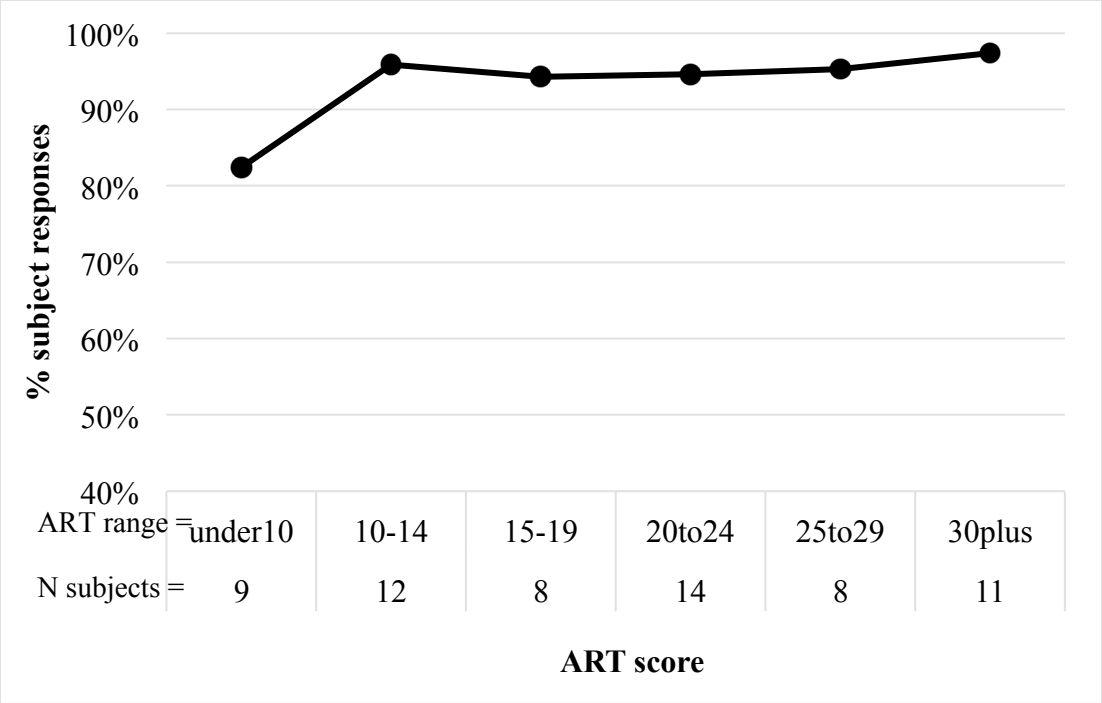

Figure 6. Plot of individual participants' average rate of selecting the subject character, as a function of their ART score (print exposure).

Again we examined the correlations between ART scores and questions about reading/internet. ART score was correlated with time spent reading books ($r = .28$, $p = .03$), and with enjoyment of reading ($r = -0.53$[23], $p < .0001$), but not with the time spent browsing internet sites ($r = -0.03$, $p = 0.79$), or time spent listening to books ($r = .17$, $p = .18$).

We also examined reaction times, in a mixed effects linear regression model, following the same procedure as in Experiments 1 and 2.[24] The centered control predictors retained in the final model were trial order (faster for later items), whether the subject option was on top (faster), and whether the verb was wants (slower) vs. needs (faster). We added centered predictors for subject response, ART; an additional model confirmed that the interaction between

[23] This is negative because our responses were 1 = love it; 2 = like it; 3 = neither like nor dislike it, and 4 = not very enthusiastic about it, so this shows that higher ART scores correlate with more enjoyment of readi

[24] The slope for trial order by items was estimated to be zero and removed.

subject response and ART was not significant. There were random effects for both participants and items, and maximal random slopes, except that trial order by item was estimated to be zero by the model and removed. There was a nonsignificant trend for responses to be faster for people with higher ART scores (β = -0.002 (0.001), t = -1.71, p = .102). Responses were also marginally faster for subject vs. nonsubject responses (β = -0.05 (0.025), t = -1.98, p = .061).

Discussion

The most important finding was that print exposure predicted pronoun comprehension, even in the absence of gaze cues. This strongly suggests that the effect is related to the use of the subject bias, and not related to the use of gaze cues. Moreover, this effect persisted even though the overall rate of selecting the subject was quite high, around 94%.

The fact that we found an ART effect in Experiment 3 suggests something about how exposure affects pronoun processing. We hypothesized that reading exposure might provide the right kind of context for people to develop the subject bias for pronoun comprehension. What exactly could people be learning through reading? One possibility is that reading exposure determines whether people acquire the subject bias at all. However, this explanation seems unlikely, given that most subjects exhibited some degree of a subject bias in Experiment 3. In fact, even in the presence of gaze cues in Experiments 1 and 2, most participants picked the subject character more than 50% of the time (55/56 subjects in Experiment 2, and 55/61 subjects in Experiment 1). Thus, the difference across subjects is not whether they exhibit a subject bias during pronoun comprehension at all, but rather how consistently they do so. Low-ART subjects may simply follow a subject assignment strategy less consistently than high-ART subjects. In

addition, they appear to be more susceptible to following other sources of information, as they were in Experiments 1 and 2.

The results from Experiment 3 also suggest that individual differences in pronoun comprehension can sometimes correlate with other measures of linguistic skill, namely the SAT reading test (although we did not find this in Experiment 2). Even though the subject bias is not a question of correct vs. incorrect understanding, more skilled readers in this experiment were also more consistent in spoken language comprehension, specifically in their assignment of pronouns to the subject. However, neither print exposure nor pronoun comprehension were strongly related to socioeconomic variables.

General Discussion

Across three experiments, our finding of greatest interest was that individual differences in print exposure predicted the tendency for people to choose the linguistically salient first-mentioned character as the referent of the pronoun. We could not explain this effect in terms of working memory (Experiment 1), theory of mind (Experiment 1), or socioeconomic status (Experiments 2 and 3). We did find that print exposure correlated with other measures of language skill.

These findings are important for several reasons. First, they extend recent work showing that experience affects lexical and grammatical processing, and show that similar effects apply at the discourse level. Other studies have reported individual difference effects on pronoun comprehension (e.g., Daneman & Carpenter, 1980; Oakhill & Yuill, 1986), but critically these studies have examined processing during reading tasks. That is, we already knew that reading skill affects specific reading processes, including pronoun comprehension and drawing

inferences. However, we did not know that print exposure affects pronoun comprehension in **spoken** language, and with stimuli using very simple sentence structures. The current study is the first to show that it does. This suggests that individual differences affect basic language processing mechanisms, and not just reading ability.

Second, our three experiments found stronger effects of reading exposure than of reading skill. Although these are sometimes correlated, the effects of exposure were more consistent than those of skill in our studies. This suggests that exposure to print materials affects the discourse comprehension strategies that people use, even in spoken comprehension. It is unlikely that reading exposure is the only thing that matters; individuals also experience differences in spoken language experience. Thus, the current findings should not be taken as evidence that print exposure matters more than spoken language experience. Instead, it is the first definitive demonstration that exposure affects specific strategies during pronoun comprehension.

This then raises the question of what exactly people learn through exposure that affects pronoun comprehension. There are several possibilities consistent with our findings. First, it could be general attention to the linguistic context. Understanding pronouns requires knowing that the two utterances are related to each other. Thus, participation in connected discourses may lead people to pay attention to the linguistic context and seek connections between utterances. A second, more specific possibility is that print exposure could lead to stronger representations of information status cues. Multiple aspects of linguistic form reflect information status categories, such as pronominalization, word order, and syntactic structure (e.g., Birner & Ward, 1998; Chafe, 1976, 1994, Lambrecht, 1994). Language exposure offers the opportunity to observe the relevance of information status, and strengthen attention to information status itself. A third possibility is that exposure may even have more specific effects, offering users evidence about

the specific patterns of reference, such as which referents are more probable. Subjects tend to get re-mentioned frequently in discourse, which means that they form parts of referential chains. This helps strengthen the generalization that the grammatical subject position signals a high likelihood of re-mention. At a broader level, the predictability of subjects also provides evidence to language learners that subjects are topical. This view of print exposure is consistent with current models of pronoun comprehension, which depend on calculations about both the probability of reference, and the probability of pronoun usage (Arnold, 1998; Kehler & Rohde, 2013). Exposure provides the input necessary to strengthen knowledge of these probabilities.

This study also ruled out other possible effects of language experience. A priori we considered the possibility that reading might strengthen the use of rare structures, as it does for syntactic and lexical processing. We found that it did not, and instead people with more experience tended to follow the dominant pattern. However, it is premature to conclude that experience has different effects on pronoun processing than syntactic processing. Notably, our task presented people with ambiguous pronouns, which were never disambiguated. In studies where exposure supports the interpretation of rare syntactic structures, there is typically disambiguating information (e.g., Fine & Jaeger, 2013; Wells et al., 2009). It may be that exposure helps the development of multiple skills related to reference understanding. Here we found that it was associated with both faster and more consistent assignment of the pronoun to the subject, which is the dominant pattern. Future work is necessary to determine whether it also affects people's ability to integrate disambiguating information.

One question that arises from our findings is whether our results are about exposure generally, or reading in particular. Questions about reading in Experiments 2 and 3 suggest that print exposure is specifically related to reading practices. However, we cannot rule out the

possibility that the effect of language exposure is general, because reading exposure may correlate with other types of spoken language exposure. That is, people who read a lot might also participate in other complex linguistic activities, which also provide the exposure needed to learn about frequent discourse patterns. This might occur, for example, if both spoken and written language experience were driven by socioeconomic status, and participation in educational and occupational activities involving language. Yet socioeconomic status in our sample did not predict pronoun comprehension, which weakens this possibility.

On the other hand, our findings are also consistent with the idea that written language provides a critical type of experience for learning discourse patterns. Written language tends to be coherent and thematically organized. It also includes more complex language. If pronoun comprehension is influenced by referential probability, exposure to complex language may be especially important for learning that subjects tend to be mentioned again more than nonsubjects. In contrast, the recency bias could be learned from any discourse in which the same entity is mentioned twice (e.g., *X got dressed. X went out*). However, learning the subject bias requires learning that the subject is relatively more likely to be mentioned than other entities in the sentence. This requires exposure to discourses with multiple referents per sentence, which are probably more frequent in written than spoken language. Written language is also decontextualized. This forces readers to look for contextual information in the linguistic context, since social gestures and other physical information is irrelevant.

While the above explanations are speculative, we can also rule out other possibilities. Experiment 3 demonstrated that the effect of print exposure is not restricted to situations in which social cues are present. We also were able to rule out the possibility that ART scores are a proxy for working memory, theory of mind, or socio-economic status. Together, these findings

suggest that print exposure affects pronoun comprehension strategies. One possibility might be that print exposure strengthens knowledge of referential frequencies, which in turn increases the speed or confidence with which participants use referential frequency to guide pronoun interpretation.

There are several open questions about the scope of this effect. We used a version of the ART that included authors mostly known for fiction. Other work has used the ART to test knowledge of both fiction and nonfiction writers (Mar et al., 2006). For example, Mumper and Gerrig (2017) did a meta-analysis that included 15 such studies. They found that both reading of fiction and nonfiction was associated with higher theory of mind and empathy scores, although the effect size for theory of mind was somewhat stronger for fiction reading. For pronoun comprehension, an open question is whether reference patterns differ for fiction and nonfiction, and whether fine-grained differences in reading preferences correlate with pronoun comprehension strategies.

It may initially seem surprising that we found no effect of working memory, which has been shown to modulate pronoun comprehension in previous tasks (Daneman & Carpenter, 1980; Nieuwland & van Berkum, 2006; van Rij et al., 2011; van Rij, van Rijn, & Hendriks, 2013). Yet our pronoun task differed from previous tasks in that it was relatively simple, tested only off-line responses, and required few discourse inferences. Participants knew that the questions following the story could query three or four pieces of information: 1) who wanted/needed the object, 2) what the object was, 3) which side each participant was on, or, in Experiment 3, 4) what they were doing. Thus, they only needed to keep track of these pieces of information. This information was readily available in the context, did not require semantic integration, and may not have imposed a severe working memory challenge. In addition, we only

examined ambiguous pronouns and final interpretation, rather than on-line processing (unlike Nieuwland & van Berkum, 2006). Thus, our findings suggest that working memory capacity does not necessarily play a large role in the final, off-line interpretation of simple spoken utterances, but it is not inconsistent with other evidence that working memory can influence on-line processing or the use of more complex discourse information.

In sum, the current project provides a critical new piece of evidence for understanding how individuals vary in their mechanisms of discourse understanding. Our findings clearly establish that language exposure, in particular print exposure, affects processing in healthy adults. This project sets the stage for studies examining how written and spoken language input influence the development of discourse processing strategies in healthy and disordered populations.

Acknowledgements

The authors thank the following assistants for testing participants: Natasha Vazquez, Bryan Smith, Michaela Neely, Anita Simha, Grant Huffman, and Megan Fullarton (Exp. 1); Margarita Rodriguez, Leah Daniels, Harvey Liu, Saradine Pierre, Darith Klibanow, and Sophia Barsanti. (Exps. 2 and 3). This work was partially supported by NSF grant 1651000 to J. Arnold. Iris Strangmann's work on Experiment 1 was supported through the University of Groningen Fund for Excellent Students as well as the Marco Polo Fund from the University of Groningen.

References

Acheson, D. J., Wells, J. B., & MacDonald, M. C. (2008). New and updated tests of print exposure and reading abilities in college students. *Behavior Research Methods*, *40*(, 278–289. doi:10.3758/brm.40.1.278

Almor, A. (1999). Noun-phrase anaphora and focus: The informational load hypothesis. *Psychological Review*, *106*, 748–765.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Arnold, J. E. (1998). *Reference form and discourse patterns* (Doctoral Dissertation). Stanford University.

Arnold, J. E. (2001). The effect of thematic roles on pronoun use and frequency of reference Continuation. *Discourse Processes*, *31*, 137–162.

Arnold, J. E. (2010). How speakers refer: the role of accessibility. *Language and Linguistics Compass*, *4*, 187–203. doi:10.1111/j.1749-818X.2010.00193.x

Arnold, J. E., Castro, L., Zerkle, S., & Rao, L. (In preparation). Print exposure predicts pronoun comprehension strategies in children. Ms., University of North Carolina.

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.

Baddeley, A. (1992). Working Memory. *Science*, *255*, 556–559.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" test revised version: A study with normal adults, and adults with asperger syndrome or high-functioning autism. *Journal of Child Psychology and Psychiatry*, *42*, 241–251.

Birner, B. J., Ward, G. (1998). Information status and noncanonical word order in English. Amsterdam/Philadelphia: John Benjamins.

Brennan, S. E., Friedman, M. W., & Pollard, C. J. (1987). A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting on Association for Computational Linguistics* (pp. 155–162). Stroudsburg, PA, USA: Association for Computational Linguistics.

Brown, C. (1983). Topic continuity in written English narrative. In T. Givón, (Ed.), *Topic Continuity* (pp. 313-342). Amsterdam: John Benjamins.

Calvo, M. G. (2001). Working memory and inferences: Evidence from eye fixations during reading. *Memory*, *9,* 365–381

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, *22*, 77–94.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. N. Li (Ed.), *Subject and Topic*, (pp. 25–56). New York: Academic Press.

Chafe, W. (1994). *Discourse, consciousness, and time*. Chicago: Chicago University Press.

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Review of Anthropology*, *16*, 383–407.

Chambers, C. G., & Smyth, R. (1998). Structural parallelism and discourse coherence: A Test of Centering theory. *Journal of Memory and Language*, *39*, 593–608.

Cook-Gumperz, J., & Gumperz, J. (1981). From oral to written culture: The transition to literacy. *Writing: The Nature, Development and Teaching of Written Communication*, *1*, 89–109.

Cowles, H. W., Walenski, M., & Kluender, R. (2007). Linguistic and cognitive prominence in anaphor resolution: topic, contrastive focus and pronouns. *Topoi*, *26*, 3–18.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Farmer, T. A., Fine, A. B., Misyak, J. B., & Christiansen, M. H. (2017). Reading span task performance, linguistic experience, and the processing of unexpected syntactic events. *The Quarterly Journal of Experimental Psychology*, *70*, 413-433. doi:10.1080/17470218.2015.1131310

Farmer, T. A., Monaghan, P., Misyak, J. B., & Christiansen, M. H. (2011). Phonological typicality influences sentence processing in predictive contexts: reply to Staub, Grant, Clifton, and Rayner (2009). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1318–1325.

Fine, A. B., & Jaeger, F. T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science*, *37*(3), 578–591.

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PLOS ONE*, *8*, e77661.

Foraker, S., & McElree, B. (2007). The role of prominence in pronoun resolution: Active versus passive representations. *Journal of Memory and Language*, *56*(3), 357–383.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science, 336*, 998. doi: 10.1126/science.1218633

Garnham, A. (2001). *Mental Models and the Interpretation of Anaphora*. Psychology Press.

Gernsbacher, M. A., & Hargreaves, D. J. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*, 699–717.

Goodrich Smith, W., & Hudson Kam, C. L. (2012). Pointing to “her”: The effect of co-speech gesture on pronoun resolution. *Language and Cognition*, *4*, 75–98.

Gordon, P. C., Grosz, B. J., & Gilliom, L. A. (1993). Pronouns, names, and the centering of attention in discourse. *Cognitive Science*, *17*, 311–347.

Gordon, P.C., Moore, M., Choi, W., Hoedemaker, R.S., & Lowder, M. (under revision). Individual differences in reading: Separable effects of knowledge and processing skill.

Grober, E. H., Beardsley, W., & Caramazza, A. (1978). Parallel function strategy in pronoun assignment. *Cognition*, *6*, 117–133.

Gundel, J. K., Hedberg, N., & Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, *69*(2), 274–307.

Hanna, J. E., & Brennan, S. E. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language*, *57*, 596–615.

Hartshorne, J.K., O'Donnell, T.J., & Tenenbaum, J.B. (2015). The causes and consequences explicit in verbs. *Language, Cognition, and Neuroscience, 30*, 716-734.

Hendriks, P., & Spenader, J. (2006). When production precedes comprehension: An optimization approach to the acquisition of pronouns. *Language Acquisition*, *13*, 319–348.

James, A. N., Fraundorf, S. H., Lee, E. K., & Watson, D. G. (under review). Individual differences in syntactic processing: Evidence from verb bias, relative clause extraction, and attachment preferences.

Kaiser, E. (2011). Focusing on pronouns: Consequences of subjecthood, pronominalisation, and contrastive focus. *Language and Cognitive Processes*, *26*, 1625–1666.

Kameyama, M. (1996). Indefeasible semantics and defeasible pragmatics. In M. Kanazawa, C. Pinon, & H. de Swart (Eds.), *Quantifiers, deduction and context* (pp. 111–138). Stanford, CA: CSLI.

Kehler, A. (2007). Rethinking the SMASH approach to pronoun interpretation. In J. K. Gundel & N. Hedberg (Eds.), *Reference: Interdisciplinary Perspectives* (pp. 95–122). Oxford: Oxford University Press.

Kehler, A., & Rohde, H. (2013). A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics*, *39*, 1–37.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*, 1–44.

Kidd, D. C., & Castano, E. (2013). Reading literary fiction improves theory of mind. *Science*, *342*, 377–380.

Lambrecht, K. (1994). *Information structure and sentence form: Topic, focus, and the mental representation of discourse referents*. New York, NY: Cambridge University Press.

Linderholm, T. (2002). Predictive inference generation as a function of working memory capacity and causal text constraints. *Discourse Processes*, *34*(3), 259–280.

MacDonald, M. C. (1994). Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes*, *9*(2), 157–201.

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*, 1–16.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*, 35–54.

Mar, R. A., Oatley, K., Hirsh, J., de la Paz, J., & Peterson, J. B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability,

and the simulation of fictional social worlds. *Journal of Research in Personality*, 40, 694–712. http://dx.doi.org/10.1016/j.jrp.2005.08.002

Martin-Chang, S. L., & Gould, O. N. (2008). Revisiting print exposure: exploring differential links to vocabulary, comprehension and reading rate. *Journal of Research in Reading*, *31*, 273–284.

Mol, S. E., & Bus, A. G. (2011). To read or not to read: A meta-analysis of print exposure from infancy to early adulthood. *Psychological Bulletin*, *137*, 267–296.

Montag, J. L., & MacDonald, M. C. (2015). Text exposure predicts spoken production of complex sentences in eight and twelve year old children and adults. *Journal of Experimental Psychology: General*, *144*, 447–468.

Moore, M., & Gordon, P. C. (2014). Reading ability and print exposure: item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*(4), 1095–1109.

Mumper, M. L., & Gerrig, R. J. (2017). Leisure reading and social cognition: A meta-analysis. *Psychology Of Aesthetics, Creativity, And The Arts*, *11*(1), 109-120. doi:10.1037/aca0000089

Nappa, R., & Arnold, J. E. (2014). The road to understanding is paved with the speaker's intentions: Cues to the speaker's attention and intentions affect pronoun comprehension. *Cognitive Psychology*, *70*, 58–81.

Nieuwland, M. S., & van Berkum, J. (2006). Individual differences and contextual bias in pronoun resolution: Evidence from ERPs. *Brain Research*, *1118*, 155–167.

Oakhill, J., & Yuill, N. (1986). Pronoun resolution in skilled and less-skilled comprehenders: effects of memory load and inferential complexity. *Language and Speech*, *29*, 25–37.

Panero, M. E., Weisberg, D. S., Black, J., Goldstein, T. R., Barnes, J. L., Brownell, H., & Winner, E. (2016). Does reading a single passage of literary fiction really improve theory of mind? An attempt at replication. *Journal of Personality and Social Psychology*, *111*(5), e46-e54. doi:10.1037/pspa0000064

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment*, *28*, 164–171.

Samur, D., et al. (in press). Does a single session of reading literary fiction prime enhanced mentalising performance? Four replication experiments of Kidd and Castano (2013). *Cognition & Emotion*. doi: 10.1080/02699931.2017.1279591

Shipley, W. C. (1940). A self-administered scale for measuring intellectual impairment and deterioration. *Journal of Psychology*, 9, 371-377.

Solomon, R. L., & Howes, D. H. (1951). Word frequency, personal values, and visual duration thresholds. *Psychological Review*, *58*, 256–270.

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*, 402–433.

Stevenson, R. J., Crawley, R. A., & Kleinman, D. (1994). Thematic roles, focus and the representation of events. *Language and Cognitive Processes*, *9*, 519–548.

Tanenhaus, M. K., & Trueswell, J. C. (1995). Sentence comprehension. In J. L. Miller & P. D. Eimas (Eds.), *Speech, language, and communication* (pp. 217–262). San Diego, CA, US: Academic Press.

Tannen, D. (1979). *Processes and Consequences of Conversational Style* (Doctoral Dissertation). University of California, Berkeley.

Tannen, D. (1980). Implications of the oral/literate continuum for cross-cultural communication. In J. Alatis (Ed.), *Current issues in bilingual education* (pp. 326–347). Washington, DC: Georgetown University Press.

Tannen, D. (1987). Conversational style. In H. Dechert & M. Raupach (Eds.), *Psycholinguistic models of production* (pp. 251-67). Norwood, NJ: Ablex.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior Research Methods*, *37*, 498–505.

van Rij, J., van Rijn, H., & Hendriks, P. (2011). WM Load Influences the Interpretation of Referring Expressions. In F. Keller & D. Reitter (Eds.), *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics* (pp. 67–75). Stroudsburg, PA, USA: Association for Computational Linguistics.

van Rij, J., van Rijn, H., & Hendriks, P. (2013). How WM load influences linguistic processing in adults: A computational model of pronoun interpretation in discourse. *Topics in Cognitive Science*, *5*, 564–580.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128.