Studies in Second Language Acquisition (2023), **45**, 853–881
doi:10.1017/S0272263122000468

RESEARCH ARTICLE

Individual differences modulate sensitivity to implicit causality bias in both native and nonnative processing

Tingting Wang[1,3]* and Alison Gabriele[2,3]

[1]Department of East Asian Languages & Cultures, University of Kansas, Lawrence, KS, USA; [2]Department of Linguistics, University of Kansas, Lawrence, KS, USA; [3]Second Language Acquisition Laboratory, Department of Linguistics, Dole Human Development Center, Lawrence, KS, USA
*Corresponding author: Email: tingtingwang@ku.edu

(Received 06 September 2021; Revised 24 September 2022; Accepted 14 October 2022)

Abstract

The question of whether L2 learners can use discourse cues online during pronoun resolution remains debated in the field. We examine one factor that has been argued to impact pronoun resolution in native speakers, implicit causality (IC) bias, a property related to certain verbs in which one of verb's arguments are considered to be the cause of an action. We investigate whether individual differences modulate sensitivity to IC bias in both native English speakers and Chinese-speaking learners of English, examining whether variability is similarly explained in the two populations. Results from a sentence completion task and a self-paced reading (SPR) task show similar sensitivity to IC bias in both groups; reading times on the SPR task were also modulated by working memory and vocabulary knowledge. The findings suggest that L2 learners are successful in using discourse-level cues during processing and that variability is qualitatively similar in both learners and natives.

Introduction

A core goal of theories of second language (L2) acquisition and processing is to better understand the domains in which L2 acquisition is successful as compared to the domains that seem to present persistent difficulty, even for L2 learners who have otherwise achieved high levels of proficiency. The Interface Hypothesis proposes that linguistic properties that lie at the interface of syntax and discourse and require the integration of multiple sources of linguistic information present persistent challenges in L2 acquisition (Sorace, 2011; Sorace & Filiaci, 2006). Referential dependencies, which occur when two linguistic expressions are interpreted as referring to the same individual, have been highlighted as one such domain that may be vulnerable for L2 learners (Sorace, 2011). Referential dependencies require that a link be established between a new referent, such as a pronoun, and a representation of that entity in the discourse model (e.g., Clark et al., 1977), a mental representation of the characters and the events mentioned in the discourse (e.g., Johnson-Laird, 1983). The interpretation of a

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

pronoun requires the integration of linguistic information, such as gender, which can be used to resolve the pronoun in (1) if the subject is *Paul*, as well as information relevant to the discourse, such as which entity in the discourse is most accessible or salient (e.g., Ariel, 1990, 2001). Discourse information may be particularly important when morphosyntactic properties cannot be used to unambiguously select an antecedent as in (1) if the subject is *Mary*.

(1) Paul/Mary admired Ruby because **she** baked with vegan ingredients.
(2) Ruby visited Frances because **she** had time off of work.
(3) Ruby impressed Mary because **she** baked with vegan ingredients.

One factor that has been argued to impact the accessibility or salience of a discourse entity is syntactic prominence, as it is more likely for a pronoun to refer to an antecedent in subject position as opposed to object position (e.g., Brennan, 1995; see Arnold, 2010 for a review). Thus in (2), *she* is more likely to be interpreted as *Ruby*. Recent research has suggested that L2 learners can use information such as syntactic prominence to guide pronoun resolution in contexts similar to (2) (e.g., Contemori et al., 2019; Cunnings et al., 2017), but have more difficulty in more complex contexts such as when two discourse entities are introduced into the discourse with similar prominence as in coordinate noun phrases (*Ruby and Frances*) (e.g., Contemori et al., 2019; Roberts et al., 2008).

A discourse entity's semantic role also plays a role in determining how salient or accessible it is in the discourse. In (1), when *Mary* is the subject and thus there are two potential gender-matching antecedents for the pronoun *she*, the pronoun is more likely to be interpreted as referring to *Ruby* as there is a bias for the pronoun to refer to the entity who is most likely to have been the cause of the event (Brown & Fish, 1983; Garvey & Caramazza, 1974). In (3), when the verb changes to *impressed*, the cause or stimulus argument is now in subject position, and the most likely antecedent for the pronoun is also the subject, *Ruby*. This phenomenon, which has been referred to as implicit causality (IC), has recently become a topic of interest in the L2 literature because of its potential to shed light on the kind of information that L2 learners can use in processing (Cheng & Almor, 2017, 2019; Contemori & Dussias, 2019; Kim & Grüter, 2021; Liu & Nichols, 2010). Recent theoretical accounts of IC bias argue that verbs such as *admire* and *impress* trigger an expectation for an explanation of the stimulus argument of the verb, or the cause of the event, and the coreference bias is related to the expectation that the specific cause will be explained (Bott & Solstad, 2014, 2021). On this account, IC bias is related to both verbal properties and discourse coherence, in that an expectation for an explanation is generated (Kehler et al., 2008). Results in the L2 literature thus far are mixed with respect to whether or not L2 learners have been observed to use this information similarly to native speakers. Cheng and Almor (2017, 2019), based on their studies of Chinese-speaking learners of English, proposed that L2 learners may be more likely to exhibit a subject bias during pronoun resolution, in line with the syntactic prominence factor described in the preceding text, as opposed to integrating the IC bias of the verb (but see Liu & Nicol, 2010). These results are important because they suggest that L2 learners may weight information in the discourse differently than native speakers, giving prominence in pronoun resolution to factors such as subjecthood, as opposed to also integrating information related to the verb and discourse coherence. However, there is also reason to believe that characteristics of the learners may play a role as the Chinese-speaking learners of English in a study by Liu and Nichol (2010) did show native-like sensitivity to IC bias. Although

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

both studies tested advanced learners of English, the learners tested in Cheng and Almor's studies were tested in China while the learners tested in Liu and Nichols's study were students in the United States. It is possible that individual differences in language experience and exposure could play a role in determining whether L2 learners can successfully use these cues.

There is support for this idea in research on IC bias in native (L1) speakers. A recent study by Johnson and Arnold (2021) reported that L1 English speakers' language experience modulated their use of IC bias in pronoun resolution. Language experience was measured by the Author Recognition Task (ART; Acheson et al., 2008; Moore & Gordon, 2015; Stanovich & West, 1989), which is designed to measure participants' exposure to print materials by providing participants with a list of names that includes both real author names and "foil" names and asking participants to decide whether or not a given name represents an actual author. The experiment involved listening to stories and making judgments about reference. Results showed that as ART scores increased, English speakers were more likely to interpret the pronoun as referring to the implicit cause encoded by the verb. These results suggest that certain individuals can use discourse cues to resolve pronouns more successfully than others and that there is variability even in native speakers. As we will review in the next section, several other papers have shown that other individual-level characteristics such as working memory (Koornneef et al., 2016) and reading abilities (Long & De Ley, 2000) modulate the use of IC bias in native speakers. Despite these findings, L2 studies have generally not explored the role of individual-level characteristics in modulating the use of IC bias in pronoun resolution outside of considering the role of L2 proficiency, which has not been found to be a significant factor (e.g., Kim & Grüter, 2021). Thus, a key open question is what explains the variability in previous studies on implicit causality in L2. Are L2 learners truly restricted in integrating discourse cues such as implicit causality to resolve pronouns successfully or does success depend on the individual-level characteristics of the learner? Do the same individual differences modulate variability in native speakers and L2 learners? The present study examines these questions by examining English native speakers and Chinese-speaking learners of English in two experiments, one which included an offline sentence completion task, and the second, which included both a self-paced reading task and two measures of individual differences. The study aims to shed light on the kind of information that L2 learners can use during referential processing and more importantly, the specific conditions under which they are successful, thus extending what we know about the factors that may facilitate pronoun resolution in L2 learners (Sorace, 2011).

Use of Implicit Causality Bias in Native Speakers

IC was defined by Garvey and Caramazza (1974) as the information encoded in verbs that implicitly attributes the cause of the action to one of the antecedents mentioned earlier in the sentence. For instance, when reading or listening to a sentence with an NP1-biased verb such as *frighten* as in (4a) in the following text, readers tend to think that the upcoming context likely relates to *David.* In contrast, when the verb is changed to the NP2-biased verb *fear* as in (4b), a possible continuation could be one that attributes the cause of the event in the main clause to *Mary.*

(4) a. David frightened Mary (because)….
 b. David feared Mary (because)….

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

The IC bias is made clear in a discourse with a causal relation that can be explicitly marked by discourse connectors such as *because*, but the bias holds even if the causal relation is implicit (Kehler et al., 2008).

Researchers have proposed various accounts in terms of the underlying mechanism of IC bias. Some researchers argue that IC bias stems from individuals' world knowledge about the causes and effects of various events (Corrigan, 2002; Pickering & Majid, 2007). Other accounts argue that IC bias is related to the thematic roles that specific verbs assign to their arguments (Brown & Fish, 1983; Crinean & Garnham, 2006; Hartshorne et al., 2015). A recent account proposed by Bott and Solstad (2014) (see also Bott & Solstad, 2021) incorporates a semantic account, but crucially considers discourse coherence as well (Kehler et al., 2008). Bott and Solstad (2021) argue that the semantic properties of IC verbs trigger expectations or preferences for specific kinds of explanations, which are intimately linked to one of the verb's arguments. For example, in (4a), the phrase *David frightened Mary* triggers an expectation that the stimulus or cause of the event will be explained because the specific cause has not been provided in this context. This has implications for reference because if the specific cause of the "frighten" event is indeed specified in the following clause, it is more likely that David will be rementioned or interpreted as the antecedent of a subsequent pronoun. In (4b), in which the NP2 *Mary* is the stimulus argument, it is more likely that an explanation of the event will mention her. As one piece of evidence for this account, Bott and Solstad (2021) show that IC bias can be manipulated by specifying an explanation of the cause in the context. Because of this evidence that shows an important role for discourse coherence and the broader context, we refer to IC bias in this article as a discourse cue.

Successful use of IC bias has generally been reported by L1 speakers in various languages by using both offline and online methods (Cozijn et al., 2011; De La Fuente, 2015: Spanish; Featherstone & Sturt, 2010: English; Holler & Suckow, 2016: German; Koornneef & Van Berkum, 2006: Dutch; Pyykkönen & Järvikivi, 2009: Finnish; Rigalleau et al., 2004: French). Recent studies employing time-sensitive methods (e.g., self-paced reading, eye-tracking) have generally reported that sensitivity to IC bias can be activated early in the sentence, at or before the pronoun (Featherstone & Sturt, 2010; Greene & McKoon, 1995; Koornneef & Van Berkum, 2006; Long & De Ley, 2000; McKoon et al., 1993; Pyykkönen & Järvikivi, 2009).

However, several studies have also demonstrated that native speakers' ability to use IC bias may be modulated by individual differences (Johnson & Arnold, 2021; Koornneef et al., 2016; Koornneef & Mulders, 2017; Long & De Ley, 2000; Van Berkum et al., 2013). An early study by Long and De Ley (2000) showed a relationship between reading abilities and use of IC bias. Their study used three probe tasks where participants read sentences as in (5) that either contained a match or a mismatch between the noun phrase associated with the verb bias (*Evette* for the NP2 verb *envy*) and the noun phrase most likely to be associated with the pronoun in the subordinate clause based on the semantic context. Participants were asked to read the sentences, and judge whether they had seen a probe word (e.g., *Sherry*/*Evette*) earlier in the sentence. They predicted faster responses to the probe words in the match condition as compared to the mismatch condition.

(5) a. Sherry envied Evette all the time because **she** had a fast car. (match)
    b. Sherry envied Evette all the time because **she** had no money. (mismatch)

Long and De Ley (2000) also categorized participants as skilled/less skilled readers on the basis of performance on the Nelson–Denny Reading Test, arguing that computing

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

causal inferences, as is required in these contexts, is complex, and may require increased resources that may be afforded by higher level reading skills. Indeed, the results showed that it was only skilled readers who showed an early effect of IC bias (right after the pronoun), responding faster to names that matched with the verb bias, but this effect was limited to NP2 verbs. Less skilled readers only showed sensitivity to IC bias at the end of the sentence. They concluded that the ability to use the causal information encoded by the verb depends on characteristics of both the stimuli and of the reader. With respect to why reading skills may facilitate the use of IC bias, Long and De Ley (2000) propose that skilled readers may have better word recognition skills or a higher level of resources available to generate causal inferences. The study by Johnson and Arnold (2021), which was discussed in the preceding text, makes a related claim. In their study, individuals with higher print exposure were better able to use IC bias to make predictions about the likelihood of remention. They propose that individuals with higher print exposure may be better at using the discourse information available to generate causal inferences. They also consider the possibility that higher print exposure may be related to a higher quality of input that enables individuals to learn which patterns of reference are more likely to occur in a discourse.

The relationship between working memory and the use of IC bias was explored by Koornneef et al. (2016) in an eye-tracking experiment. Following Long and De Ley (2000), they predicted an early use of IC bias for individuals with higher working memory (as measured by a digit span task). In addition, in one block of the experiment, they added a secondary task (storing and recalling a sequence of digits) to examine if the use of IC bias would be modulated by the amount of processing resources available. The results showed a complex pattern. In the condition without the secondary task, an IC effect emerged at the pronoun only for individuals with higher working memory; a similar effect emerged three words after the pronoun as well. In contrast, in the condition with the secondary task, at the region three words after the pronoun, the opposite pattern emerged: It was the lower-span readers who showed a more pronounced IC effect. Thus, surprisingly, lower-span readers were more likely to use IC bias in pronoun resolution under a processing burden; Koornneef et al. (2016) interpret this finding as evidence against Long and De Ley's (2000) argument that the use of IC bias is related to the amount of resources available. Koornneef et al. (2016) propose that the differences between high- and low-span individuals might be explained in terms of reading strategies: higher-span readers may, under normal circumstances, rely on a proactive reading strategy, generating expectations about reference, while lower-span readers do not generate these expectations, and wait instead for bottom-up information. In contrast, under a processing burden, higher-span individuals may become more conservative, waiting for bottom-up information, while lower-span readers take a "risky" approach to compensate for the lack of available resources.

Use of Implicit Causality Bias by L2 Learners

In the L2 literature, results are mixed with respect to whether or not learners have been observed to use IC bias successfully in pronoun resolution. Cheng and Almor (2017, 2019), based on their studies of Chinese-speaking learners of English, proposed that L2 learners may be more likely to exhibit a subject or first mention bias in assigning reference to a pronoun as opposed to integrating the IC bias of the verb. Their experiments utilized sentence-completion tasks including sentence fragments that consisted of two same-gendered names, with a verb that carried either NP1 or NP2

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

bias, the connector *because*, and an ambiguous pronoun (e.g., *Paul liked Alan because he…*). Cheng and Almor (2017) showed that both native English speakers and L2 learners displayed a strong NP1 bias for NP1 verbs, but learners displayed a weaker NP2 bias by showing more NP1 choices for NP2 verbs than native speakers. In a follow-up study, Cheng and Almor (2019) found that learners showed a weaker association between pronouns and NP1 referents in contexts containing NP1 verbs, but a stronger association between pronouns and NP1 referents following NP2 verbs.

In contrast, Liu and Nicol (2010) also observed similar performance for L1 Chinese L2 English learners and English natives. Their study used a self-paced reading task where Chinese learners of English ($n = 41$) and native English speakers ($n = 41$) read sentences that contained either a match (plausible condition) or a mismatch (implausible condition) between the verb bias and the gender of the pronoun (e.g., *The mother$_i$ amused the father because he$_{*i}$ told funny jokes at dinner*). The results show that both L1 speakers and L2 learners showed significant reading time slowdowns in the implausible condition as compared to the plausible condition for both NP1 and NP2 verbs.

More recent studies have used visual-world eye-tracking to examine whether IC bias can be used to proactively predict which antecedent is most likely to be referred to based on the discourse context. A study by Kim and Grüter (2021) showed an early use of IC bias by L1 English speakers starting before the pronoun, whereas L1 Korean L2 English learners showed a weaker effect of IC bias starting after the pronoun offset. Thus, the L2 learners showed a similar pattern, but differed from the natives with respect to the strength and timing of the effect. However, a study by Contemori and Dussias (2019) that used a similar design did not find group differences between L1 English speakers and highly proficient Spanish-English bilinguals, a result that they attributed to the advanced proficiency level of the bilinguals, who were exposed to English at an early age and who had lived in the United States for many years. Similar to the point we raised in the "Introduction," this is another example that suggests that the background characteristics of the L2 learners may play an important role in determining whether or not L2 learners can use IC bias similarly to native speakers. While the role of proficiency in the L2 is a natural candidate for potentially explaining the variability in findings, analyses by Kim and Grüter (2021) did not find proficiency to be a significant factor. Thus, the present study takes a different approach, examining the role of individual differences that have been found to be significant in the native processing literature and testing whether variability in the use of IC bias is similarly explained in native speakers and L2 learners.

Current Study

The main goal of the present study is to examine the role of individual differences in the use of IC bias during pronoun resolution by both native speakers and learners to examine whether processing in the two populations is qualitatively similar or different. The study will address two main questions. First, we investigate whether L2 learners show sensitivity to IC bias in resolving pronouns offline and online. Second, we examine whether individual differences modulate successful use of IC bias for both natives and L2 learners. The study includes two experiments: Experiment 1 uses an offline sentence completion task and Experiment 2 uses a self-paced reading task as well as two individual difference measures.

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Experiment 1: Sentence Completion Task

The first experiment, which used a sentence-completion task, had two main goals. First, we wished to confirm that the experimental sentences that we designed for the self-paced reading study to be reported in Experiment 2 did indeed show the intended verb biases. Second, this task also allowed us to examine whether L2 learners have similar biases as native speakers when given unlimited time and when asked to focus somewhat more explicitly on reference.

Participants

We included 67 L1 English speakers (44 females, mean age = 19.6, range: 18–28) in the United States and 69 Chinese-speaking learners of English (59 females, mean age = 22.0, range: 18–31) who were university students in China. All learners reported that they started to learn English as a second language in a school setting from age 5 onward (*M* = 10, range: 5–14). The LexTALE, a lexical decision task consisting of 60 items made up of both real and nonce words, was used as a proficiency measure (Lemhöfer & Broersma, 2012). The learners' mean score of 67.5/100 (*SD* = 11.57, range: 47.8–97.5), places them at an intermediate to advanced proficiency level. All participants completed the task online using *Qualtrics*. Native speakers were offered extra credit in a course and learners were offered monetary compensation for their participation.

Materials

We used a sentence-completion task, adopting the three-sentence design used in Koornneef and Van Berkum (2006). In this design, two different-gendered names are initially introduced in the first sentence by a conjoined noun phrase (*Lindsey and Brad*) and are then referred to by a plural pronominal *they* in the second sentence (see 6/7). By doing so, we hope to attenuate a potential first-mention bias by introducing the two potential antecedents into the discourse with equal prominence (Gordon et al., 1999). The third sentence is the target sentence, where two blanks in the main clause and a blank in the subordinate clause were provided for completion. We selected 36 sets of unique name pairs stereotypically associated with either male or female pronouns. Participants were instructed to use the two names to fill in the first two blanks, and to provide a natural ending to the story starting with either *he* or *she*, by typing the completed version of the third sentence into a text box.

(6) Example stimuli of NP1 verbs
Lindsey and Brad were working at a homeless shelter. They chatted seriously about how to better help the homeless.
________ fascinated _______ because (he/she) ____________________________

(7) Example stimuli of NP2 verbs
William and Brittany were living together. They both liked the house to be clean.
______ appreciated _______ because (he/she) ____________________________

We selected 18 NP1 and 18 NP2 English verbs with strong IC bias in both English and Chinese based on three previous studies (Cheng, 2016; Ferstl et al., 2011; Hartshorne et al., 2013). Details on the verb selection in the study are provided in *Supplementary*

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Materials-Verb Selection (https://osf.io/b3cqp/). Two lists were prepared: In one list, half the lead-in sentences started with a male name while the other half started with a female name, and the order of the two names were switched in the second list. In addition to 36 target trials, five "catch" trials that instructed participants to type the sentence "I love dogs and cats" into a response box randomly appeared during the experiment, serving as a test of whether participants were indeed attending to the test items, as the experiment was administered remotely.

Procedure

Participants completed a language background questionnaire first, followed by the sentence-completion task. Finally, L2 learners also completed the LexTALE.

Data analysis

Continuations were coded as either NP1 or NP2 choice based on whether the pronoun matched with the first or second antecedent in gender. Continuations that contained the wrong verb or did not include a third-person singular gendered pronoun (e.g., *it*, *they*) were eliminated from the analysis (L1: 4.6%, L2: 2.9%). Following previous studies (Cheng & Almor, 2017, 2019; Contemori et al., 2019; Kim, 2019), we used NP1 choice as the dependent variable for the analyses. Mixed-effects logistic regression (Baayen, 2008; Jaeger, 2008) was used in examining the proportion of NP1 choice in NP1 and NP2 stories. The model included *Group* (Natives, Learners) and *VerbType* (NP1, NP2) as fixed effects (sum-coded), and *Participant* and *Item* as random effects. The model also contained the maximal random effects structure by including a by-participant slope for *VerbType*. The by-item slope for *Group* was excluded to solve a model convergence problem. Data analysis was conducted using the lme4 package in R (Bates et al., 2015). Degrees of freedom for the *t*-values and *p*-values in the mixed-effects models were computed using the R package lmerTest (Kuznetsova et al. 2017).

Results

The model revealed a significant main effect of VerbType ($b = 2.15$, SE = .12, $z = 17.39$, $p < .001$), with a significant interaction between VerbType and Group ($b = -.65$, SE = .05, $z = -13.04$, $p < .001$). Follow-up analysis was conducted by using the "emmeans" package in R (Lenth, 2021) in comparing the NP1 choice between native speakers and learners in stories with NP1 verbs and NP2 verbs (the Holm method was chosen to control family-wise error rate). Results showed that in the NP1 stories, learners chose NP1 significantly less than native speakers ($b = -1.56$, SE = .15, $z = -10.48$, $p < .001$), while in the NP2 stories, learners chose NP1 significantly more than native speakers ($b = 1.05$, SE = .13, $z = 7.91$, $p < .001$) (Figure 1).

The offline results suggest that L2 learners are sensitive to the IC bias encoded by the verb when resolving pronouns. However, their sensitivity was not as strong as native speakers. The results do not suggest that L2 learners were relying on a heuristic such as a general subject/first-mention bias (Cheng & Almor, 2017) because learners showed a weaker bias for both the NP1 and the NP2 biased verbs, as opposed to showing an overall NP1 bias in both verb types.

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

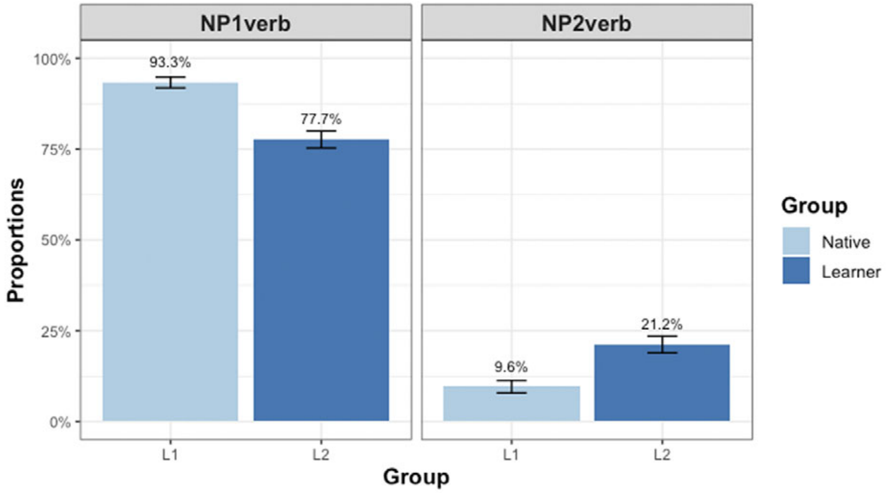

Figure 1. Mean proportion of NP1 choice by native speakers and learners.

Experiment 2: Examining the Online Use of IC Bias by L1 and L2 Speakers of English

Participants

We recruited 40 native English speakers (9 males, mean age = 19.3, age range: 18–23) and 39 Chinese learners of English[1] (9 males, mean age = 24.7, age range: 19–41) from a university in the United States. We assessed proficiency using the LexTALE (Lemhöfer & Broersma, 2012), and the University of Michigan Listening Comprehension Test, which targets various aspects of English grammar. Based on the proficiency scores (LexTALE: 60.8/100; Michigan: 89/100), learners were characterized as high intermediate-advanced learners. All participants were offered course credit or monetary compensation for their participation. Learners' English-learning background information as well as the results from the two proficiency tests are summarized in Table 1.

Materials

Self-Paced Reading Task

The main task utilized a word-by-word noncumulative moving window self-paced reading paradigm. Our design was inspired by Koornneef and Van Berkum (2006) who examined implicit causality using self-paced reading in Dutch. Examples from the NP1 and NP2 conditions are given in Table 2. We used the same 36 three-sentence stories from Experiment 1, but following Koornneef and Van Berkum (2006), the target sentence was manipulated such that in the Consistent Condition (a/c), there was a match between the gender of the pronoun (*he*) and the stereotypical gender of the noun phrase associated with the IC bias (NP1/NP2), and in the Inconsistent Condition (b/d), there was a mismatch between the gender of the pronoun and the gender of the noun

[1] One L2 participant was excluded from the analysis due to low accuracy in answering the comprehension questions related to the experiment.

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Table 1. Descriptive statistics for L2 learners' background information.

	Onset of English study (age)	Years of English study	Age of arrival	Years living in the US
M (*SD*)	9.5 (2.2)	13.2 (3.5)	22.2 (4.8)	1.5 (2.2)
	Michigan Test	**LexTALE**		
M (*SD*)	89/100 (5.4)	60.8 (9.3)		

Table 2. Examples of the stimuli in experiment 2

Verb type	Lead-in	Target sentence
NP1 story	Nick and Lisa were working as lawyers. They cared about helping in the community.	a. *Consistent (gender-match)* **Nick**$_{i1}$ inspired$_2$ Lisa$_3$ because$_4$ **he**$_{i5}$ had$_6$ been$_7$ trying$_8$ hard to help people who couldn't afford legal fees. b. *Inconsistent (gender-mismatch)* **Lisa**$_{i1}$ inspired$_2$ Nick$_3$ because$_4$ **he**$_{i*5}$ had$_6$ been$_7$ trying$_8$ hard to help people but she went above and beyond.
NP2 story	Bob and Lily were both working for the same law firm. They knew that only one person could become a partner.	c. *Consistent (gender-match)* Lily$_1$ disliked$_2$ **Bob**$_{i3}$ because$_4$ **he**$_{i5}$ had$_6$ been$_7$ using$_8$ the company's resources for his own personal use. d. *Inconsistent (gender-mismatch)* Bob$_1$ disliked$_2$ **Lily**$_{i3}$ because$_4$ **he**$_{i*5}$ had$_6$ been$_7$ using$_8$ his personal time to correct all of her mistakes.

phrase associated with the antecedent bias of the verb. We predicted a reading time slowdown at the pronoun in the Inconsistent Conditions as compared to the Consistent Conditions.

Regions 1–4 included two names with opposite genders, the IC verb, and the discourse connector *because*. Region 5, the critical region, included the pronoun. Following Koornneef and Van Berkum (2006) and Featherstone and Sturt (2010), we used *he* as the pronoun in the target sentences to avoid possible reading time differences of pronouns in different genders. Regions 6–8 were the spillover regions, which were kept identical between the two conditions in the same set. The target sentences in both the Consistent and Inconsistent conditions wrapped up in a plausible way. In half the target sentences, the order of the names in the lead-in sentence matched that of the target sentence while in the other half they didn't. We also included 36 filler stories using verbs with no known IC biases. Fillers were similar to the target stories, but the target sentences either did not include pronouns or included different pronouns (*they*, *she*). For all stories, lead-in sentences were presented in full, one sentence at a time, while the target sentence was read by participants word by word. Participants also read four practice trials.

Measures of Individual Differences

Following Koornneef et al. (2016), we selected a nonverbal measure of working memory, the counting span task (Case et al., 1982). During the task, participants were presented with an array of target shapes (dark blue circles) and distractors (light green circles). On each trial, they were asked to count the number of target shapes out loud (in their native language) and then repeat the total number; the experimenter then entered that number, which triggered the next trial. After a series of 2–6 trials, a series of boxes appeared on the screen prompting the participants to recall the total number of target shapes from each previous trial in order. Instructions were given in the participants' L1.

We also included the Peabody Picture Vocabulary Test 4th edition (PPVT-4; Dunn & Dunn, 2007) as a measure of English vocabulary knowledge. While Long and De Ley (2000) used a standardized reading test, we did not have the time in our experimental session to include a comprehensive reading assessment. Johnson and Arnold (2021) used the ART as a measure of language experience/print exposure, but the English version of the task has not been found to be appropriate for L2 learners (see McCarron & Kuperman, 2021). Because better vocabulary knowledge has been shown to be a predictor of both reading skills (Braze et al., 2007) and language comprehension in native speakers (Van Dyke et al., 2014), we felt a vocabulary test may be a reasonable replacement and would be appropriate for both L2 learners and native speakers. The PPVT-4 is widely used with English natives and has been reported to be a suitable measurement for assessing vocabulary size in more advanced L2 learners (Goriot et al., 2018). Participants were shown four pictures and were instructed to choose the correct picture that matched the target word that the experimenter said out loud. Each set contained 12 trials, and the sets were ordered with increasing difficulty. For each participant, a raw score was calculated by subtracting the number of errors from the number of completed items. We did not use the standardized score as these age-based scores are normed for native speakers. Instead, we used raw PPVT scores for both natives and learners.

Lexical Tasks

L2 learners also completed two additional tasks. First, they completed a task to examine the IC bias for the Chinese translations of the English verbs used in the experiments. The Chinese task was modeled on Hartshorne et al. (2015). We calculated learners' mean bias scores of the Chinese counterpart of the English IC verbs. If learners chose the antecedent that aligned with the expected verb bias, the choice was given a score of 1. If learners chose the antecedent that didn't align with the verb bias, the choice was given a score of 0. Learners generally displayed a strong bias in the expected direction for Chinese NP1 verbs ($M = 0.84/1$, $SD = 0.15$, range: 0.44–1.00) and for Chinese NP2 verbs ($M = 0.96/1$, $SD = 0.07$, range: 0.67–1.00). This suggests that the IC bias is similar in Chinese and English and, thus, crosslinguistic differences in verb bias should not impact processing.

We also administered a translation task to make sure that learners were familiar with the target verbs. We asked participants to translate the main clause of all 36 target sentences (e.g., *Lily disliked Bob*) into Chinese. Correctly translated items were scored 1 while incorrectly translated ones were given 0. Learners had an average score of 0.90/1 ($SD = 0.15$, range: 0.42–1.00) for NP1 items, and 0.94/1 ($SD = 0.1$, range: 0.76–1.00) for NP2 items, suggesting that they were familiar with the lexical items.

Procedure

All participants first provided informed consent and completed a language background questionnaire. L2 learners were also shown two vocabulary lists prior to beginning the

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

experiment to ensure familiarity with the lexical items in the stories. One list contained all the critical verbs and a selection of nouns/phrases, along with Chinese translations. A second list contained all the English names used in the main task to ensure familiarity.

Both natives and L2 learners completed the main self-paced reading (SPR) task and the working memory task on a computer in a quiet lab. Learners also completed the Michigan Listening Comprehension Test right after the working memory task. Then, the vocabulary task (PPVT-4) was administered to both groups by the experimenter. Lastly, learners completed the two additional lexical tasks on a laptop using *Qualtrics.* Participants received $10 per hour for their participation; the session lasted approximately one hour for the natives, and two hours for the learners.

Predictions

Target sentences for each condition are listed in Table 3. If participants use IC bias, they should slow down at the pronoun (e.g., *he*) in the Inconsistent Conditions as compared to the Consistent Conditions in both NP1/NP2 sentences. However, if, as proposed by Cheng and Almor (2017, 2019), learners resolve the pronoun in favor of the first-mentioned/subject antecedent, they would slow down at the pronoun in the Inconsistent Condition for NP1 sentences, but they would show the reverse pattern for NP2 sentences. Specifically, for NP2, they would slow down at the pronoun in the Consistent Condition as the subject in that condition mismatches in gender with the pronoun.

For individual differences, based on the previous studies described in the preceding text, we predict a positive modulation of slowdowns at the pronoun in the Inconsistent Condition compared to the Consistent Condition based on participants' working memory abilities and vocabulary scores.

Statistical Analysis

Data Preprocessing for the SPR Task

Only the target trials for which the comprehension question was correctly answered were included in the analysis (mean accuracy: L1: 96%; L2: 93%). In line with Nicklin and Plonsky (2020), we excluded reading times falling outside of the range of 150 ms and 2000 ms, which resulted in the removal of 2.9% of native speakers' data and 1.0% of learners' data. We then log-transformed the reading times to repair the skewness of the data distribution. *SD* boundaries were not used in our data trimming process since it

Table 3. Target sentences for NP1/NP2 sentences

Verb type	Condition	Example
NP1	Consistent (gender-match)	**Nick**$_i$ inspired Lisa because **he**$_i$ had been trying hard to help people who couldn't afford legal fees.
	Inconsistent (gender-mismatch)	**Lisa**$_i$ inspired Nick because **he**$_{i*}$ had been trying hard to help people but she went above and beyond.
NP2	Consistent (gender-match)	Lily disliked **Bob**$_i$ because **he**$_i$ had been using the company's resources for his own personal use.
	Inconsistent (gender-mismatch)	Bob disliked **Lily**$_i$ because **he**$_{i*}$ had been using his personal time to correct all of her mistakes.

has been argued that log-transformation circumvents the need for such method, which has the potential of altering potentially legitimate data (Nicklin & Plonsky, 2020).

Data Preprocessing for the Individual Differences Measures

We used mean accuracy in the counting span task, and raw scores in the PPVT-4 task. The two individual differences (ID) scores were z-transformed when entered into the statistical models.

Statistical Analysis

Log-transformed reading times (logRTs) were statistically analyzed by fitting linear mixed-effects models in R (R Core Team, 2021) using lme4 (Bates et al., 2015). Separate analyses were conducted for the pronoun (pro), and the spillover regions (pro+1, pro+2, pro+3). We included three spillover regions in line with Koornneef and Van Berkum (2006), who analyzed five words after the pronoun and revealed an effect in a SPR experiment at two words following the pronoun. Other studies that used a similar gender-mismatch paradigm also found that IC effect emerged in the spillover regions (Liu & Nicol, 2010, SPR task: two words after the pronoun; Featherstone & Sturt, 2010, eye-tracking: one and three words after the pronoun).

In the model that examined the IC effect at the pronoun, *Consistency* (Consistent, Inconsistent), *Group* (Natives, Learners),*VerbType* (NP1, NP2), and the two ID measures: *Counting Span* and *PPVT*, as well as all possible interactions among those variables were included as fixed effects. In the model that examined the IC effect at the spillover regions, *Consistency* (Consistent, Inconsistent), *Group* (Natives, Learners), *VerbType* (NP1, NP2), *Region* (pro+1, pro+2, pro+3), and the two ID measures: *Counting Span* and *PPVT*, as well as all possible interactions among those variables were included as fixed effects. The IC effect at the precritical region (pro-1) was also examined by including *Consistency*, *Group*, and *VerbType* as fixed effects to ensure that no reading time differences emerged before the critical region; the analysis showed that no significant differences were observed. In all the models, categorical factors were sum coded and the two ID factors were z-transformed. All models also contained the maximal random effects structure, including by-participant slopes for Consistency, VerbType, and Region (only for models of the spillover regions), and a by-item slope for Group.[2] Data analysis was conducted using the lme4 package in R (Bates et al., 2015). Degrees of freedom for the *t*-values and *p*-values in the mixed-effects models were computed using the R package lmerTest (Kuznetsova et al., 2017).

Descriptive Statistics of ID Measures

Descriptive statistics for the ID measures are shown in Table 4. Pearson correlation tests showed that the two ID measures are not significantly correlated (r (76) = .2, p = –.15). The results of t-tests also showed that there was no statistically significant difference in the Counting Span scores between the two groups (t (73.6) = –1.64, p = .10), but native speakers' raw PPVT scores were significantly higher than learners' (t (46.2) = 12.14, p < .001).

[2] Random slopes were removed if they were either estimated to be 0 or the model did not converge.

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Table 4. Descriptive statistics for the individual difference measures

	Counting span (%)		PPVT (raw)	
	M (*SD*)	Range	*M* (*SD*)	Range
Natives	65.72 (12.56)	37.78-94.22	204.8 (8.75)	179–216
Learners	70.74 (14.32)	39.78-94.67	154.6 (24.04)	99–197

Results

Figure 2 illustrates the mean RTs for native speakers and learners in items combining NP1 and NP2 verbs. Table 5 presents the mean RTs and standard deviations of each region. At the group level, the descriptive data show that effects are very subtle and at the pronoun, learners' mean reading times are in the opposite of the predicted pattern. However, our inclusion of the individual difference measures allows us to examine variability in both groups, investigating whether the effects, which are not robust at the group level, depend on working memory or vocabulary knowledge. In what follows, we present the statistical results for the pronoun and the spillover regions (pro+1, pro+2, pro+3) in detail, including measures of individual differences.[3]

Critical Region (Pronoun)

At the pronoun, we were interested in whether participants showed sensitivity to IC bias and whether the sensitivity would vary by individual's performance on the Counting Span and PPVT-4. The model showed no significant main effect of Consistency but showed a significant interaction between Consistency and Counting Span ($b = -0.007$, SE $= 0.003$, $t = -2.212$, $p = .030$) (Table 6).

To understand the significant Consistency by Counting Span interaction, we plotted the relation between individual's Counting Span scores and their reading times of the pronoun at the Consistent and Inconsistent Conditions in Figure 3. The figures suggest that both native speakers and learners with higher working memory showed larger reading time slowdowns at the pronoun in the Inconsistent Condition compared to the Consistent Condition when combining both NP1 and NP2 sentences. The lack of a three-way interaction between Consistency, Counting Span, and Group suggests that working memory modulates sensitivity to the IC bias similarly in native speakers and learners.[4]

[3] A reviewer pointed out that learners seem to show longer reading times at the verb in the Consistent Condition as compared to the Inconsistent Condition, which is unexpected. We think it is possible that the unexpected pattern at the verb might be related to the ordering of the male and female names in the clause with the IC verb, with certain events being perceived as more or less likely to have a female/male agent (e.g., *Lily disliked Bob* vs. *Bob disliked Lily*). It is important to point out that this reading time slowdown did not spill over to the discourse connector *because*, which is the precritical region, and thus any effects observed at the pronoun or spillover regions cannot be attributed to differences that emerged earlier in the sentence.

[4] A reviewer pointed out that some of the participants with lower working memory scores, particularly in the Learner group, show an opposite effect in which reading times are slower in the Consistent condition compared to the Inconsistent condition. A similar tendency was observed by Koornneef et al. (2016). It is important to point out that there is not a three-way interaction between Consistency, Counting Span, and Group and, thus, while the effect is visually more present in the learners, we don't have statistical evidence for group differences. We acknowledge that this effect was unexpected and that a future study with a larger

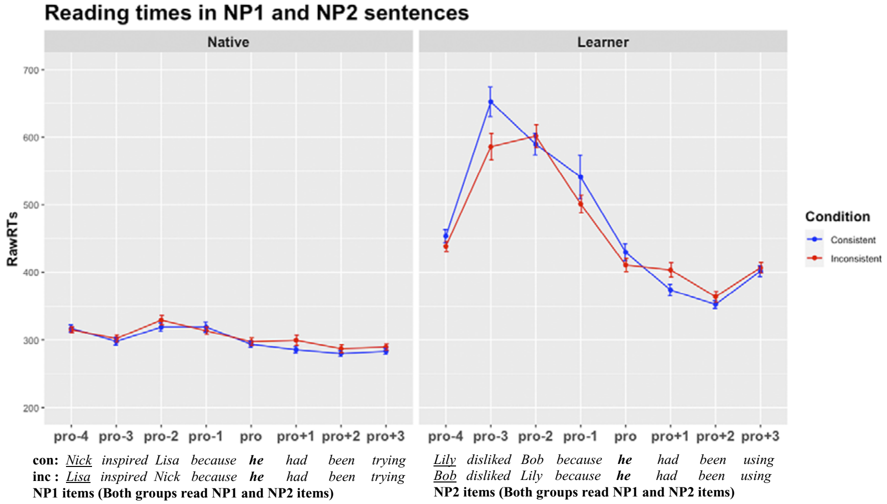

Figure 2. Mean RTs for native speakers and learners in consistent and inconsistent items (NP1 and NP items combined).

Table 5. Descriptive statistics for raw reading times and standard deviations in the pronoun and the spillover regions (pro+1, pro+2, pro+3)

Group	Condition	Pronoun	Pro+1	Pro+2	Pro+3
Natives	Consistent	294(118)	286(128)	280(117)	283(120)
	Inconsistent	297(148)	300(199)	287(154)	290(119)
Learners	Consistent	430(320)	374(213)	353(147)	402(207)
	Inconsistent	411(246)	404(269)	364(200)	407(195)

Spillover Regions (pro+1, pro+2, pro+3)

We also analyzed the spillover regions to capture any late effects (e.g., Featherstone & Sturt, 2010; Koornneef & Van Berkum, 2006; Liu & Nicol, 2010). The results showed a significant main effect of Consistency (b = –.005, SE = .002, t = –2.207, p = .031), with a significant two-way interaction between Consistency and PPVT (b = –.007, SE = .002, t = –3.180, p = .002). There were also two significant three-way interactions that involved the variable Consistency: a significant interaction between Consistency, Group, and VerbType (b = –.004, SE = .001, t = –2.808, p = .005); and a significant interaction between Consistency, Group, and PPVT (b = –.005, SE = .002, t = –2.038, p = .045). Lastly, there were also three significant four-way interactions that involved the variable Consistency: a significant interaction between Consistency, Group, VerbType, and PPVT (b = –.004, SE = .002, t = –2.503, p = .012); a significant interaction between Consistency, Group, VerbType and CS (b = .003, SE = .002, t = 2.108, p = .035); and a significant interaction between Consistency, VerbType, PPVT, and CS (b = –.004, SE = .002, t = –2.230, p = .026) (Table 7). The results indicate that the effect of Consistency, and the modulation of this effect by the two ID measures differed depending on the participant

sample size would allow us to better investigate whether individuals with lower working memory show a qualitatively different pattern than individuals with higher working memory.

Table 6. Results of the mixed-effects regression at the critical region (pronoun) (N = 78)

Predictors	*b*	*SE*	*df*	*t*	*p*	
(Intercept)	2.509	0.010	77.490	239.157	0.002	***
Consistency1	0.002	0.003	69.327	0.510	0.612	
Group1	0.071	0.010	70.084	7.033	0.000	***
VerbType1	0.008	0.004	34.033	2.002	0.053	.
PPVT	−0.009	0.010	70.184	−0.889	0.377	
CS	0.003	0.010	70.104	0.324	0.747	
Consistency1:Group1	0.003	0.003	69.287	1.026	0.309	
Consistency1:VerbType1	−0.003	0.003	2452.968	−0.960	0.337	
Group1:VerbType1	0.002	0.003	2453.220	0.844	0.399	
Consistency1:PPVT	−0.001	0.003	71.473	−0.338	0.736	
Group1:PPVT	0.008	0.010	70.185	0.798	0.428	
VerbType1:PPVT	0.002	0.003	2452.222	0.673	0.501	
Consistency1:CS	−0.007	0.003	71.102	−2.212	0.030	*
Group1:CS	−0.006	0.010	70.107	−0.610	0.544	
VerbType1:CS	0.001	0.003	2449.510	0.394	0.693	
PPVT:CS	0.008	0.011	70.267	0.739	0.462	
Consistency1:Group1:VerbType	−0.001	0.003	2451.093	−0.203	0.839	
Consistency1:Group1:PPVT	0.001	0.003	70.879	0.435	0.665	
Consistency1:VerbType1:PPVT	−0.004	0.003	2483.112	−1.432	0.152	
Group1:VerbType1:PPVT	0.000	0.003	2452.836	−0.113	0.910	
Consistency1:Group1:CS	−0.002	0.003	69.924	−0.601	0.550	
Consistency1:VerbType1:CS	0.002	0.003	2469.681	0.888	0.375	
Group1:VerbType1:CS	−0.003	0.003	2451.419	−1.252	0.211	
Consistency1:PPVT:CS	0.003	0.003	71.292	0.806	0.423	
Group1:PPVT:CS	−0.009	0.011	70.269	−0.807	0.423	
VerbType1:PPVT:CS	−0.002	0.003	2450.676	−0.622	0.534	
Consistency1:Group1:VerbType1:PPVT	−0.001	0.003	2476.466	−0.520	0.603	
Consistency1:Group1:VerbType:CS	0.000	0.003	2468.610	0.112	0.911	
Consistency1:Group1:PPVT:CS	−0.001	0.003	71.268	−0.271	0.787	
Consistency1:VerbType1:PPVT:CS	−0.005	0.003	2457.387	−1.566	0.117	
Group1:VerbType1:PPVT:CS	0.001	0.003	2451.344	0.435	0.664	
Consistency1:Group1:VerbType1:PPVT:CS	−0.005	0.003	2457.043	−1.691	0.091	.

Note: The random slopes of VerbType and Group were removed due to the singularity warning.
Formula: lmer (logRT ~ Consistency * Group * VerbType * PPVT * CS + (1+ Consistency|Participant) + (1|Item))

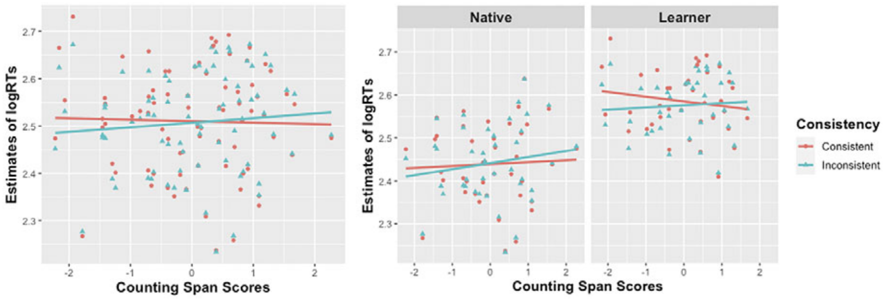

Figure 3. Relationship between counting span scores and reading times at the pronoun in consistent (red line) and inconsistent (blue line) conditions when combining two participant groups (left) and splitting the data by group (right).

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Table 7. Results of the mixed-effects regression model at the spillover regions (pro+1, pro+2, pro+3) (N = 78)

Predictors	*b*	*SE*	*df*	*t*	*p*	
(Intercept)	2.486	0.011	76.080	229.909	0.000	***
Consistency1	–0.005	0.002	69.480	–2.207	0.031	*
Group1	0.061	0.011	71.830	5.711	0.000	***
Region1	0.002	0.002	7573.000	1.021	0.307	
Region2	–0.013	0.002	7573.000	–6.164	0.000	***
VerbType1	0.008	0.003	49.300	2.679	0.010	**
PPVT	–0.012	0.011	70.070	–1.129	0.263	
CS	–0.002	0.011	70.050	–0.201	0.841	
Group1:Region2	–0.009	0.002	7573.000	–4.113	0.000	***
Consistency1:PPVT	–0.007	0.002	71.850	–3.180	0.002	**
Consistency1:Group1:VerbType1	–0.004	0.001	7622.000	–2.808	0.005	**
Consistency1:Group1:PPVT	–0.005	0.002	71.550	–2.038	0.045	*
Group1:VerbType1:PPVT	0.004	0.002	69.620	ed1.887	0.063	.
Consistency1:Group1:VerbType1:PPVT	–0.004	0.002	7119.000	–2.503	0.012	*
Consistency1:Group1:VerbType1:CS	0.003	0.002	6697.000	2.108	0.035	*
Consistency1:VerbType1:PPVT:CS	–0.004	0.002	7635.000	–2.230	0.026	*

Note: The random slope of Region was removed due to the singularity warning.
Formula: lmer (logRT ~ Consistency * Group * Region * VerbType * PPVT * CS (1+ Consistency + VerbType|Participant) + (1+Group|Item))

group and the verb type. To further examine these interactions, we conducted follow-up analyses splitting the data by *VerbType*. Due to length considerations, in the tables that follow, we only report the results for the fixed effects and all significant interactions. See Supplementary Materials: Full Model Output for the comprehensive set of results (https://osf.io/b3cqp/).

Spillover Regions: NP1 items

Results of the model that only contained NP1 items didn't show any significant effect of Consistency but showed a significant two-way interaction between Consistency and PPVT ($b = -.008$, $SE = .003$, $t = -2.676$, $p = .009$), and a significant three-way interaction between Consistency, Group, and PPVT ($b = -.008$, $SE = .003$, $t = -2.643$, $p = .01$) (Table 8). This suggests that the Consistency effect at the spillover regions in NP1 items was dependent on participants' PPVT-4 scores, and the pattern differed between natives and learners.

To further investigate the Consistency by Group by PPVT interaction, we first conducted a follow-up analysis comparing the Consistency effect between the two groups using the "emmeans" package in R (Lenth, 2021). The results showed that only learners showed a marginal reading time slowdown at the spillover regions of the NP1 items in the Inconsistent Condition compared to the Consistent Condition ($b = .018$, $SE = .009$, $t = 2.043$, $p = .082$) (Table 9).

Next, we plotted the relation between individual's PPVT-4 scores and their reading times in the two conditions (Figure 4). The figure shows that while learners with higher PPVT-4 scores showed a larger Consistency effect than learners with lower PPVT-4 scores; native speakers' sensitivity to the IC bias was not modulated by their PPVT-4 scores. Overall, analyses of the NP1 items at the spillover regions showed different patterns between native speakers and learners: Native speakers didn't show any

Table 8. Results of the mixed-effects regression model at the spillover regions for NP1 items (N = 78)

Predictors	*b*	*SE*	*df*	*t*	*p*	
(Intercept)	2.494	0.012	77.790	214.841	0.000	***
Consistency1	−0.004	0.003	69.020	−1.394	0.168	
Region1	0.003	0.003	3758.000	0.825	0.409	
Region2	−0.012	0.003	3758.000	−3.881	0.000	***
Group1	0.062	0.011	70.040	5.528	0.000	***
PPVT	−0.010	0.011	70.120	−0.885	0.379	
CS	−0.002	0.011	70.060	−0.174	0.862	
Consistency1:Region1	−0.005	0.003	3758.000	−1.750	0.080	.
Region2:Group1	−0.009	0.003	3758.000	−2.778	0.006	**
Consistency1:PPVT	−0.008	0.003	71.260	−2.676	0.009	**
Region1:PPVT	0.006	0.003	3758.000	1.845	0.065	.
Consistency1:Group1:PPVT	−0.008	0.003	70.570	−2.643	0.010	*

Note: Formula: lmer (logRT ~ Consistency * Region * Group * PPVT * CS (1+ Consistency |Participant) + (1|Item))

Table 9. Results of the Follow-up Analysis of the Consistency by VerbType Interaction for Natives and L2 Learners

Contrast	Group	*b*	*SE*	*df*	*z.ratio*	*p*
Inconsistent- Consistent	Native	0.000	0.009	Inf	-0.056	0.955
Inconsistent- Consistent	Learner	0.018	0.009	Inf	2.043	0.082

Note: *p* value adjustment: holm method for two tests.

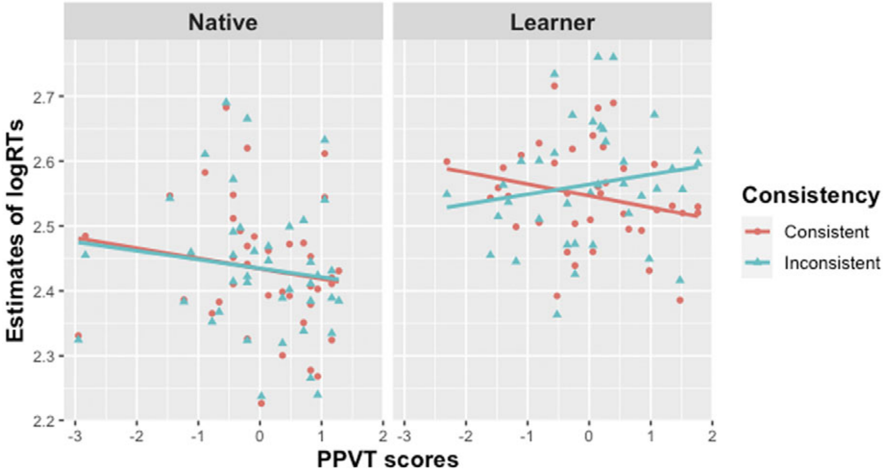

Figure 4. Relationship between PPVT scores and reading times at the spillover regions in Consistent (red line) and Inconsistent (blue line) conditions in NP1 items by the two participant groups.

sensitivity to the IC bias while learners with higher vocabulary scores showed stronger sensitivity to IC bias.[5]

[5] A reviewer pointed out that there are some learners with lower vocabulary scores who show an opposite effect in which reading times are slower in the Consistent Condition compared to the Inconsistent Condition. Following the reviewer's suggestion, we conducted post-hoc analyses examining the direction of the

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Spillover Regions: NP2 items

Results of the model that only contained NP2 items show a marginal effect of Consistency ($b = -.005$, $SE = .003$, $t = -1.912$, $p = .060$), with a significant two-way interaction between Consistency and PPVT ($b = -.006$, $SE = .003$, $t = -2.029$, $p = .046$), and a marginal three-way interaction between Consistency, PPVT, and Counting Span ($b = .006$, $SE = .003$, $t = 1.934$, $p = .057$) (Table 10). The results suggest that the effect of Consistency was dependent on individuals' PPVT-4 scores, which may be modulated differently by individuals' Counting Span scores, and the patterns were similar between natives and learners.

To further explore the Consistency by PPVT by Counting Span three-way interaction, we included Counting Span scores and PPVT-4 scores separately into the model, along with variables of Consistency, Group, Region, and all possible interactions. First, the model that included individuals' PPVT-4 scores revealed a marginal Consistency effect ($b = -.005$, $SE = .003$, $t = -1.828$, $p =. 072$), with a significant interaction between Consistency and PPVT ($b = -.006$, $SE = .003$, $t = -2.131$, $p =.036$) (Table 11).

Visualization of the relation between PPVT-4 scores and reading times indicate that both native speakers and learners with higher PPVT-4 scores showed a large effect of Consistency than those with lower PPVT-4 scores (Figure 5).[6]

Second, the model that included individuals' Counting Span scores revealed a marginal Consistency effect ($b = -.005$, $SE = .003$, $t = -1.800$, $p =. 076$) but with no significant interaction that involved Consistency or Counting Span (Table 12). The results indicate that the Consistency effect found in both natives and learners was not related to individuals' Counting Span scores, and the three-way interaction between Consistency, PPVT, and Counting Span was present due to the different modulations of the Consistency effect by individuals' PPVT scores versus their Counting Span scores.

Table 10. Results of the mixed-effects regression model at the spillover regions for NP2 items (N = 78)

Predictors	*b*	*SE*	*df*	*t*	*p*	
(Intercept)	2.477	0.011	78.640	225.959	0.000	***
Consistency1	−0.005	0.003	70.310	−1.912	0.060	.
Region1	0.002	0.003	3763.000	0.615	0.539	
Region2	−0.014	0.003	3763.000	−4.924	0.000	***
Group1	0.060	0.011	75.940	5.590	0.000	***
PPVT	−0.015	0.011	70.070	−1.356	0.180	
CS	−0.002	0.011	70.040	−0.210	0.834	
Region2:Group1	−0.009	0.003	3763.000	−3.078	0.002	**
Consistency1:PPVT	−0.006	0.003	72.100	−2.029	0.046	*
Consistency1:PPVT:CS	0.006	0.003	71.540	1.934	0.057	.

Note: Formula: lmer (logRT ~ Consistency * Region * Group * PPVT * CS (1+ Consistency|Participant) + (1+Group|Item))

Consistency by PPVT interaction in learners and found that such effect was mainly driven by learners with higher PPVT scores whose effect is in the predicted direction. The unexpected pattern in the learners with lower vocabulary scores seems to be driven by a small number of individuals. As we acknowledge in note 4, a future study with a larger sample size would allow us to better investigate whether individuals with lower vocabulary scores show a qualitatively different pattern than individuals with higher vocabulary scores.

[6] Following a reviewer's suggestion, we conducted post-hoc analyses examining the direction of the Consistency by PPVT interaction in both groups and found that the effect was driven by individuals with higher PPVT scores.

Table 11. Results of the mixed-effects regression model at the spillover regions for NP2 items including PPVT as the ID measure (N = 78)

Predictors	*b*	*SE*	*df*	*t*	*p*	
(Intercept)	2.478	0.011	82.860	227.374	0.000	***
Consistency1	−0.005	0.003	74.570	−1.828	0.072	.
Region1	0.002	0.003	3780.000	0.661	0.509	
Region2	−0.014	0.003	3780.000	−5.080	0.000	***
Group1	0.059	0.011	80.190	5.490	0.000	***
PPVT	−0.013	0.011	74.050	−1.255	0.213	
Region2:Group1	−0.009	0.003	3780.000	−3.037	0.002	**
Consistency1:PPVT	−0.006	0.003	76.070	−2.131	0.036	*
Consistency1:Region2:PPVT	−0.005	0.003	3780.000	−1.647	0.100	.

Note: The random slope of Region was removed due to the singularity warning. Formula: lmer (logRT ~ Consistency * Region * Group * PPVT (1+ Consistency + |Participant) + (1+Group|Item))

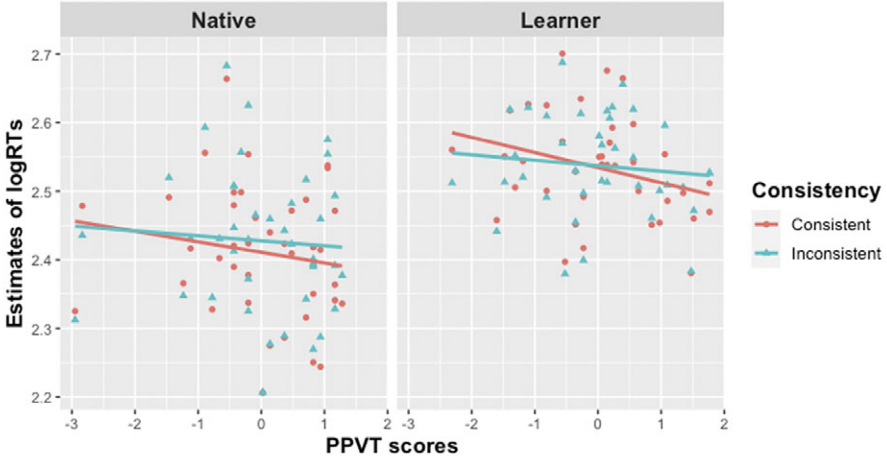

Figure 5. Relationship between PPVT scores and reading times at the spillover regions in Consistent (red line) and Inconsistent (blue line) conditions in NP2 items by the two participant groups.

Table 12. Results of the mixed-effects regression model at the spillover regions for NP2 items including Counting Span as the ID measure (N = 78)

Predictors	*b*	*SE*	*df*	*t*	*p*	
(Intercept)	2.478	0.011	82.800	226.686	0.000	***
Consistency1	−0.005	0.003	74.140	−1.800	0.076	.
Region1	0.002	0.003	3779.000	0.669	0.504	
Region2	−0.014	0.003	3779.000	−5.041	0.000	***
Group1	0.059	0.011	79.960	5.474	0.000	***
CS	−0.001	0.011	74.030	−0.102	0.919	
Region2:Group1	−0.008	0.003	3779.000	−3.004	0.003	**

Note: The random slope of Region was removed due to the singularity warning. Formula: lmer (logRT ~ Consistency * Region * Group * CS (1+ Consistency + |Participant) + (1+Group|Item))

Overall, analyses of the NP2 items at the spillover regions showed similar patterns between native speakers and learners: Participants in both groups with higher English vocabulary knowledge showed greater sensitivity to IC bias.

Summary of the SPR task

At the pronoun, a significant interaction between *Consistency* and *Counting Span* suggests that both native speakers and learners with higher working memory were more likely to show reading time slowdowns at the pronoun in the Inconsistent Condition as compared to the Consistent Condition. In the spillover regions (pro+1, pro+2, pro+3), only learners showed reading time slowdowns in the Inconsistent Condition compared to the Consistent Condition in NP1 items, an effect that was modulated by vocabulary scores. However, in NP2 items, both natives and learners showed reading time slowdowns in the Inconsistent Condition compared to the Consistent Condition, which were similarly modulated by their vocabulary knowledge. Overall, these results suggest that participants with higher working memory may have more resources available to integrate the IC bias right at the pronoun. In the spillover regions, this effect was more robust in NP1 items for learners with higher vocabulary scores, and more robust in NP2 items for both native speakers and learners with higher vocabulary scores.

A few points are worth highlighting with respect to the overall patterns that emerged. As a reviewer points out, we did not observe an overall effect of Consistency but rather an effect of Consistency that was modulated by working memory at the critical region (for both native speakers and learners) and vocabulary scores in the spillover region (for both NP1/NP2 for learners and NP2 for native speakers). A reviewer points out that the lack of an overall effect of Consistency contrasts with the previous literature. While this is true, very few studies have used word-by-word self-paced reading experiments to examine IC bias, and to our knowledge, Koornneef and Van Berkum (2006) (Experiment 1) is the only other study whose stimuli included extended discourses as opposed to using a single-sentence design (as in Liu and Nicol, 2010). Our English experiment followed the design and procedures of Koornneef and Van Berkum's (2006) Dutch experiment with just small modifications (e.g., we included 9 experimental items per condition as compared to 10 in their study and we tested n = 39 learners and n = 40 native speakers as compared to n = 24 in their study). Although Koornneef and Van Berkum (2006) observed an overall effect of Consistency, their effects were still subtle, emerging only two words after in the pronoun in the self-paced reading experiment. Thus, it is possible that our inclusion of individual difference measures in the current study allowed us to observe effects of IC bias earlier in the sentence, but only for individuals with higher working memory or higher vocabulary scores.

Secondly, although the significant interactions that emerged between Consistency and Counting Span scores and Consistency and PPVT scores are in the predicted direction, with individuals with higher working memory (at the critical region) and higher vocabulary scores (at the spillover region) showing larger effects of IC bias, there are some individuals with lower scores on the working memory and vocabulary tasks who show effects in the opposite direction. A similar tendency was observed by Koornneef et al. (2016) in their eye-tracking study that also examined the relationship between working memory and use of IC bias. Although these effects seem to be driven by few individuals, we acknowledge that a future study should continue to explore these

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

patterns and it is possible that a study with a larger sample size may show more robust qualitative differences in the processing of IC bias between individuals at the higher and lower ends of the working memory and vocabulary spectrum.

Discussion

This study investigated sensitivity to IC bias in pronoun resolution in native speakers and Chinese-speaking learners of English and explored the role of individual differences in both populations. The results suggest that in both offline and online tasks, L2 learners showed qualitatively similar patterns to native speakers. However, in the offline task, learners' preferences were weaker than the natives for both NP1 and NP2 verbs. In the SPR task, both groups showed online sensitivity to IC bias, with effects being modulated by individual's working memory and vocabulary knowledge in both groups. The results overall provide important evidence that L2 learners of English whose L1 is Chinese are sensitive to discourse-level cues such as IC bias and furthermore, that individual-level variability is similarly explained in both populations.

Different from Cheng and Almor (2017, 2019), learners in our study did not show a general subject/first-mention preference. In the offline task, not only did learners show more NP1 choices for NP2 verbs but they also showed fewer NP1 choices for NP1 verbs. Thus, our learners just showed a quantitively weaker bias than the natives, but the overall patterns were similar. We think that it is at least possible that the subject bias in Cheng and Almor's (2017) study was influenced by the single-sentence design used in the study. In sentence fragments such as *Paul liked Alan because he…*, *Paul* is introduced first in subject position, and thus, based on previous studies (Crawley et al., 1990; Frederiksen, 1981; Gernsbacher & Hargreaves, 1988), may be more prominent than the antecedent in object position. In NP1 sentences, the subject and the antecedent aligned with the IC bias of the verb point to the same discourse entity, but in NP2 sentences, the two cues point to different entities. If the subject antecedent initially carries prominence, then, in processing sentences with NP2 verbs, the discourse model needs to be updated, from the entity in subject position carrying prominence to the entity aligned with the IC bias (NP2) carrying prominence. It is possible that the learners in Cheng and Almor's studies were more likely to choose NP1 (subject) antecedents, particularly for NP2 items because they had difficulty updating the discourse model on the basis of the verb. In the present study, the three-sentence story design that we used, inspired by Koornneef and Van Berkum (2006), introduced the two possible antecedents into the discourse with equal prominence and, thus, may have weakened the prominence of the subject antecedent.

In the online task, learners' sensitivity to IC bias generally resembled that of native speakers. We will first consider the role of individual differences for both populations and then consider the findings in light of L2 theories. At the pronoun (collapsing over both NP1/NP2 items), both L2 learners and native speakers with higher working memory showed significant slowdowns at the pronoun in the Inconsistent Condition compared to the Consistent Condition. These results are in line with Koornneef et al. (2016), who showed an effect of IC bias at the pronoun only for individuals with higher working memory in the "normal" reading condition without the secondary task. Thus, we replicated their results for native speakers and extended the finding to L2 learners. An important question then is what this relationship between working memory and the use of IC bias indexes. Koornneef et al.'s original prediction was inspired by a capacity-based approach to working memory, in line with proposals such as Daneman and

Carpenter (1980) and Just and Carpenter (1992), who argue that difficulty in language comprehension is related to limitations in the processing resources available. However, Koornneef et al. (2016) argue that the results of the experiment with the secondary task do not support an account based on capacity limitations as it was individuals with lower working memory who were more likely to use IC bias when under a processing burden. As discussed earlier, Koornneef et al. propose that the patterns in their study may be explained by differences in processing strategies, with higher span individuals being more likely, under normal circumstances, to engage in proactive processing, generating expectations about which entities in the discourse are more likely to be rementioned. The relationship between working memory and anticipatory processing is a potentially interesting direction to pursue in future research in this domain, particularly with experimental paradigms that allow for a more direct test of predictive processing (e.g., Kim & Grüter, 2021; Pyykkönen & Järvkivi, 2009).

An additional consideration against capacity-based approaches is that theoretical accounts such as Lewis et al. (2006) and McElree et al. (2003) have argued that it is retrieval from memory, as opposed to storage capacity, that is critical for language comprehension (see Van Dyke & Johns, 2012 for a review). Research has shown that retrieval of the correct target in language comprehension is more difficult when similar items are stored in memory, leading to interference (e.g., Gordon et al., 2002; Van Dyke & McElree, 2006). Memory interference accounts have been applied to the domain of reference (e.g., Cunnings et al., 2014) as pronoun resolution involves retrieval of the target antecedent from memory on the basis of cues related to morphosyntax, syntax, and discourse. On memory interference accounts, individual differences in language comprehension are related to how successfully an individual uses the relevant cues to retrieve the correct target (Van Dyke & Johns, 2012). A large-scale individual differences study by Van Dyke et al. (2014), which included an extensive number of measures targeting both language skills and cognitive skills, showed that successful language comprehension (low susceptibility to interference) was best predicted by vocabulary knowledge. Working memory, which was highly correlated with many of the measures tested, did not uniquely explain any of the variance in comprehension.

In the present study, we observed a relationship between vocabulary knowledge and the use of IC bias, a finding that has, to our knowledge, not been observed previously for either native speakers or L2 learners. Our results showed that at the spillover regions, learners with higher vocabulary showed a significant IC effect for the NP1 items, while both native speakers and learners with higher vocabulary showed such a pattern in the NP2 items. It is unclear why the effect was significant for learners for both NP1 and NP2 items but was only significant for natives for NP2 items. As these effects emerged in the spillover region, it is possible that the effect was longer lasting in NP1 items for leaners with higher vocabulary knowledge. Sentences in the Inconsistent Conditions with NP1 verbs such as *Lisa inspired Nick because he…*. present contexts in which initially there is a gender mismatch between the pronoun and the name that carries the verb bias. However, all sentences wrap up in a plausible way and thus, the pronoun in the Inconsistent Condition ultimately needs to be resolved in favor of the gender-matching antecedent. This may cause processing difficulties that are prolonged in learners as compared to native speakers in certain contexts, and it is possible that it is only learners with higher vocabulary knowledge who are actively attempting to resolve the pronoun in favor of the gender-matching antecedent.

Although the finding that vocabulary knowledge modulates the use of IC bias is unique to our study, it is possible that the explanation is related to previous accounts of individual differences such as Johnson and Arnold (2021) in the domain of IC bias and

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

accounts such as Van Dyke et al. (2014) for language comprehension more broadly. Vocabulary knowledge has been shown to be related to reading skill overall (Braze et al., 2007) and it is possible that individuals with higher vocabulary/reading skills have more exposure to input that allows them to better form expectations about which discourse entities are more likely to be rementioned in certain contexts (Johnson & Arnold, 2021). It is also possible, in line with Van Dyke et al. (2014), that better vocabulary is related to higher quality lexical representations, which may facilitate the retrieval of the correct target. Successful retrieval in pronoun resolution relies on the use of a number of cues. Individuals with higher quality lexical representations may be better able to use those cues, and thus retrieval is less vulnerable to interference. Related to this issue, an anonymous reviewer asked to what extent our findings may have been influenced by the fact that we provided learners with a vocabulary list before the experiment (names used in the experiment and a selection of nouns and verbs) to ensure familiarity with the lexical items in the stimuli. While we do not believe that providing the list could have influenced their knowledge of IC bias, we acknowledge that it may have facilitated lexical access and processing overall, which could have in turn facilitated learners' use of the relevant cues in the study, such as, for example, the gender of the names. We also acknowledge that providing the learners with the list of names and not giving the list to the native speakers could have provided the learners with an advantage; a reviewer notes that the overall effect of Consistency is more robust at the spillover region for the learners than it is for native speakers. Future studies that more systematically manipulate this factor can shed light on the extent to which processing patterns are modulated by lexical familiarity.

Overall, the results of our study are in line with previous L2 studies in reporting nativelike patterns in using IC verb bias in resolving pronouns online (Contemori & Dussias, 2019; Kim & Grüter, 2021; Liu & Nicol, 2010). However, the fact that learners' biases are weaker than native speakers in the offline task and the fact that sensitivity to IC bias online is facilitated by enhanced vocabulary knowledge and working memory also suggests that the use of IC bias in pronoun resolution is quite complex. This complexity is in the spirit of the Interface Hypothesis (Sorace & Filiaci, 2006) and the RAGE hypothesis (Grüter et al., 2017) that both predict difficulty in referential processing, but the results do not support a strong version of either account, which would predict inevitable vulnerabilities in antecedent choice (Sorace & Filiaci, 2006) or a reduced ability in L2 learners to use discourse cues to generate expectations (Grüter et al., 2017). However, more recent versions of these proposals suggest a more nuanced approach (e.g., Kaan & Grüter, 2021; Sorace, 2011), and our results suggest, in line with the spirit of these recent proposals, that the goal of current research in this domain should be to better understand the specific conditions under which L2 processing is successful. The conditions likely need to account for grammatical factors (L1-L2 differences), experimental factors (nature of the task), and individual-level factors as well, which we have focused on here (see also Gabriele et al., 2021). In a comparison of native and L2 processing, while we can systematically compare the impact of experimental factors and individual-level factors, obviously the role of certain grammatical factors, such as differences between the L1 and L2, is an issue specific to the learner population being tested. For this reason, in the present study we aimed to control for L1/L2 differences and specifically examined IC verbs that carry similar biases in Chinese and English. This allowed us to focus more precisely on whether variability in the natives and the learners could be explained by similar sources. We believe that our results do indeed suggest that processing is qualitatively similar in both populations (Hopp, 2010; Kaan, 2014; McDonald, 2006). These results are an important contribution to the literature in that previous studies, such as Roberts et al. (2008), had suggested that L2 learners cannot successfully use discourse cues to resolve pronouns online, even in cases of L1-L2 similarity. We believe the examination of individual differences helps to shed light on why difficulty in this domain is often observed; the ability to use subtle discourse cues to generate implicatures as to which antecedent is in focus is modulated by an individual's language abilities, such as vocabulary knowledge, and cognitive abilities, and thus, may be beyond the range of some learners and even some native speakers. Importantly, our results suggest that the same abilities facilitate processing in both populations. We propose, in line with Kaan (2014), that variability in the two populations may be derived from similar sources, specifically individual differences related to the abilities that support the fundamental operations of language processing.

Conclusion

This study used an offline sentence completion task and an online SPR task to investigate native English speakers and L1 Chinese L2 learners of English's sensitivity to IC bias in resolving pronouns and investigated how individual differences in vocabulary and working memory modulated that sensitivity. Results showed that both native speakers and learners utilize IC bias in resolving pronouns in similar ways, with individual differences in vocabulary knowledge and working memory explaining variability in both populations. We propose, in line with Kaan (2014), that L2 processing and native processing is fundamentally similar, and that a lack of sensitivity in previous L2 studies may be due in part to not adequately capturing variability in learners.

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/S0272263122000468.

Data Availability Statement. We have no known conflicts of interest to disclose. The experiment in this article earned Open Data and Open Materials badges for transparent practices. The materials are available at https://osf.io/b3cqp/.

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. https://doi.org/10.3758/BRM.40.1.278

Ariel, M. (1990). *Accessing noun-phrase antecedents*. Routledge and Academic Press.

Ariel, M. (2001). Accessibility theory: An overview. In T. Sanders, J. Schilperoord, & W. Spooren (Eds.), *Text representation: Linguistic and psycholinguistic aspects* (pp. 29–87). John Benjamins Publishing Company. https://doi.org/10.1075/hcp.8.04ari

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

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge University Press.

Bates D., Mächler M., Bolker B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Bott, O., & Solstad, T. (2014). From verbs to discourse: A novel account of implicit causality. In B. Hemforth, B. Mertins, & C. Fabricius-Hansen (Eds.), *Psycholinguistic approaches to meaning and understanding across languages* (Vol. *44*, pp. 213–251). Studies in Theoretical Psycholinguistics. Springer. https://doi.org/10.1007/978-3-319-05675-3_9

Bott, O., & Solstad, T. (2021). Discourse expectations: Explaining the implicit causality biases of verbs. *Linguistics*, *59*, 361–416. https://doi.org/10.1515/ling-2021-0007

Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *Journal of Learning Disabilities*, *40*, 226–243. https://doi.org/10.1177/00222194070400030401

Brennan, S. E. (1995). Centering attention in discourse. *Language and Cognitive Processes*, *10*, 137–167. https://doi.org/10.1080/01690969508407091

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. *Cognition*, *14*, 237–273. https://doi.org/10.1016/0010-0277(83)90006-9

Case, R., Kurland, D. M., & Goldberg, J. (1982). Operational efficiency and the growth of short-term memory span. *Journal of Experimental Child Psychology*, *33*, 386–404. https://doi.org/10.1016/0022-0965(82)90054-6

Cheng, W. (2016). *Implicit causality and consequentiality in native and non-native coreference processing* [Doctoral dissertation, University of South Carolina]. https://scholarcommons.sc.edu/etd/3830/

Cheng, W., & Almor, A. (2017). The effect of implicit causality and consequentiality on nonnative pronoun resolution. *Applied Psycholinguistics*, *38*, 1–26. https://doi.org/10.1017/S0142716416000035

Cheng, W., & Almor, A. (2019). A Bayesian approach to establishing coreference in second language discourse: Evidence from implicit causality and consequentiality verbs. *Bilingualism: Language and Cognition*, *22*, 456–475. https://doi.org/10.1017/S136672891800055X

Clark, H. H., Haviland, S., & Freedle, R. O. (1977). *Discourse production and comprehension*. Ablex Publishing Corporation.

Contemori, C., Asiri, O., & Perea Irigoyen, E. (2019). Anaphora resolution in L2 English: An analysis of discourse complexity and cross-linguistic interference. *Studies in Second Language Acquisition*, *41*, 971–998. https://doi.org/10.1017/S0272263119000111

Contemori, C., & Dussias, P. E. (2019). Prediction at the discourse level in Spanish–English bilinguals: An eye-tracking study. *Frontiers in Psychology*, *10*, 956. https://doi.org/10.3389/fpsyg.2019.00956

Corrigan, R. (2002). The influence of evaluation and potency on perceivers' causal attributions. *European Journal of Social Psychology*, *32*, 363–382. https://doi.org/10.1002/ejsp.96

Cozijn, R., Commandeur, E., Vonk, W., & Noordman, L. G. (2011). The time course of the use of implicit causality information in the processing of pronouns: A visual world paradigm study. *Journal of Memory and Language*, *64*, 381–403. https://doi.org/10.1016/j.jml.2011.01.001

Crawley, R., Stevenson, R., & Kleinman, D. (1990). The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research*, *19*, 245–264. https://doi.org/10.1007/BF01077259

Crinean, M., & Garnham, A. (2006). Implicit causality, implicit consequentiality and semantic roles. *Language and Cognitive Processes*, *21*, 636–648. https://doi.org/10.1080/01690960500199763

Cunnings, I., Fotiadou, G., & Tsimpli, I. (2017). Anaphora resolution and reanalysis during L2 sentence processing: Evidence from the visual world paradigm. *Studies in Second Language Acquisition*, *39*, 621–652. https://doi.org/10.1017/S0272263116000292.

Cunnings, I., Patterson, C., & Felser, C. (2014). Variable binding and coreference in sentence comprehension: Evidence from eye movements. *Journal of Memory and Language*, *71*, 39–56. https://doi.org/10.1016/j.jml.2013.10.001

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466. https://doi.org/10.1016/S0022-5371(80)90312-6

De La Fuente, I. (2015). *Putting pronoun resolution in context: The role of syntax, semantics, and pragmatics in pronoun interpretation* [Doctoral dissertation, Université Paris Diderot]. https://hal.archives-ouvertes.fr/tel-01535977/

Dunn, L. M., & Dunn, D. M. (2007). PPVT-4: *Peabody picture vocabulary test.* Pearson Assessments.

Featherstone, C. R., & Sturt, P. (2010). Because there was a cause for concern: An investigation into a word-specific prediction account of the implicit-causality effect. *Quarterly Journal of Experimental Psychology*, *63*, 3–15. https://doi.org/10.1080/17470210903134344

Ferstl, E. C., Garnham, A., & Manouilidou, C. (2011). Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods*, *43*, 124–135. https://doi.org/10.3758/s13428-010-0023-2

Frederiksen, J. (1981). Understanding anaphora: Rules used by readers in assigning pronominal referents. *Discourse Processes*, *4*, 323–347. https://doi.org/10.1080/01638538109544525

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Gabriele, A., Alemán Bañón, J., Hoffman, L., Covey, L., Rossomondo, A., & Fiorentino, R. (2021). Examining variability in the processing of agreement in novice learners: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. *47*, 1106–1140. http://doi.org/10.1037/xlm0000983

Garvey, C., & Caramazza, A. (1974). Implicit causality in verbs. *Linguistic Inquiry*, *5*, 459–464. http://www.jstor.org/stable/4177835

Gernsbacher, M.A., & Hargreaves, D. (1988). Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language*, *27*, 699–717. https://doi.org/10.1016/0749-596X(88)90016-2

Gordon, P. C., Hendrick, R., Ledoux, K., & Yang, C. L. (1999). Processing of reference and the structure of language: An analysis of complex noun phrases. *Language and Cognitive Processes*, *14*, 353–379. https://doi.org/10.1080/016909699386266

Gordon, P. C., Hendrick, R., & Levine, W. H. (2002). Memory-load interference in syntactic processing. *Psychological Science*, *13*, 425–430. https://doi.org/10.1111/1467-9280.00475

Goriot, C., Broersma, M., McQueen, J. M., Unsworth, S., & Van Hout, R. (2018). Language balance and switching ability in children acquiring English as a second language. *Journal of Experimental Child Psychology*, *173*, 168–186. https://doi.org/10.1016/j.jecp.2018.03.019

Greene, S. B., & McKoon, G. (1995). Telling something we can't know: Experimental approaches to verbs exhibiting implicit causality. *Psychological Science*, *6*, 262–270. https://doi.org/10.1111/j.1467-9280.1995.tb00509.x

Grüter, T., Rohde, H., & Schafer, A. J. (2017). Coreference and discourse coherence in L2. *Linguistic Approaches to Bilingualism*, *7*, 199–229. https://doi.org/10.1075/lab.15011.gru

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. https://doi.org/10.1080/23273798.2015.1008524

Hartshorne, J. K., Sudo, Y., & Uruwashi, M. (2013). Are implicit causality pronoun resolution biases consistent across languages and cultures? *Experimental Psychology*, *60*, 179–196. https://doi.org/10.1027/1618-3169/a000187

Holler, A., & Suckow, K. (2016). How clausal linking affects noun phrase salience in pronoun resolution. In Anke Holler & Katja Suckow (Eds.), *Empirical perspectives on anaphora resolution* (pp. 61–85). De Gruyter. https://doi.org/10.1515/9783110464108-005

Hopp, H. (2010). Ultimate attainment in L2 inflectional morphology: Performance similarities between non-native and native speakers. *Lingua*, *120*, 901–931. https://doi.org/10.1016/j.lingua.2009.06.004

Jaeger, F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446. https://doi.org/10.1016/j.jml.2007.11.007

Johnson, E., & Arnold, J. E. (2021). Individual differences in print exposure predict use of implicit causality in pronoun comprehension and referential prediction. *Frontiers in Psychology*, *12*, 672109. https://doi.org/10.3389/fpsyg.2021.672109

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, *99*, 122. https://doi.org/10.1037/0033-295X.99.1.122

Kaan, E. (2014). Predictive sentence processing in L2 and L1: What is different? *Linguistic Approaches to Bilingualism*, *4*, 257–282. https://doi.org/10.1075/lab.4.2.05kaa

Kaan, E., & Grüter, T. (2021). Prediction in second language processing and learning: Advances and directions. In E. Kaan & T. Grüter (Eds.), *Prediction in second language processing and learning* (pp. 1–24). John Benjamins.

Kehler, A., Kertz, L., Rohde, H., & Elman, J. L. (2008). Coherence and coreference revisited. *Journal of Semantics*, *25*, 1–44. https://doi.org/10.1093/jos/ffm018

Kim, H. (2019). *Cross-linguistic activation in Korean L2 learners' processing of remention bias in English* [Doctoral dissertation, University of Hawaiʻi at Mānoa]. ProQuest Dissertations.

Kim, H., & Grüter, T. (2021). Predictive processing of implicit causality in a second language: A visual-world eye-tracking study. *Studies in Second Language Acquisition*, *43*, 133–154. https://doi.org/10.1017/S0272263120000443

Koornneef, A., Dotlačil, J., van den Broek, P., & Sanders, T. (2016). The influence of linguistic and cognitive factors on the time course of verb-based implicit causality. *Quarterly Journal of Experimental Psychology*, *69*, 455–481. https://doi.org/10.1080/17470218.2015.1055282

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Koornneef, A., & Mulders, I. (2017). Can we “read” the eye-movement patterns of readers? Unraveling the relationship between reading profiles and processing strategies. *Journal of Psycholinguistic Research*, *46*, 39–56. https://doi.org/10.1007/s10936-016-9418-2

Koornneef, A. W., & Van Berkum, J. J. (2006). On the use of verb-based implicit causality in sentence comprehension: Evidence from self-paced reading and eye tracking. *Journal of Memory and Language*, *54*, 445–465. https://doi.org/10.1016/j.jml.2005.12.003

Kuznetsova A., Brockhoff, P. B., & Christensen, R. H. B. (2017). ImerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*, 1–26. https://doi.org/10.18637/jss.v082.i13

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. https://doi.org/10.3758/s13428-011-0146-0

Lenth, R. V. (2021). Emmeans: Estimated marginal means, aka least-squares means. R package version. 1.7.1-1. https://CRAN.R-project.org/package=emmeans

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*, 447–454. https://doi.org/10.1016/j.tics.2006.08.007

Liu, R., & Nicol, J. (2010). Online processing of anaphora by advanced English learners. In M. T. Prior, Y. Watanabe, & S. Lee (Eds.), *Selected proceedings of the 2008 second language research forum* (pp. 150–165). Cascadilla Proceedings Project.

Long, D. L., & De Ley, L. (2000). Implicit causality and discourse focus: The interaction of text and reader characteristics in pronoun resolution. *Journal of Memory and Language*, *42*, 545–570. https://doi.org/10.1006/jmla.1999.2695

McCarron, S. P., & Kuperman, V. (2021). Is the author recognition test a useful metric for native and non-native English speakers? An item response theory analysis. *Behavior Research Methods*, *53*, 2226–2237. https://doi.org/10.3758/s13428-021-01556-y

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*, 381–401. https://doi.org/10.1016/j.jml.2006.06.006

McElree, B., Foraker, S., & Dyer, L. (2003). Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, *48*, 67–91. https://doi.org/10.1016/S0749-596X(02)00515-6

McKoon, G., Greene, S. B., & Ratcliff, R. (1993). Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 1040–1052. https://doi.org/10.1037/0278-7393.19.5.1040

Moore, M., & Gordon, P. C. (2015). Reading ability and print exposure: Item response theory analysis of the author recognition test. *Behavior Research Methods*, *47*, 1095–1109. https://doi.org/10.3758/s13428-014-0534-3

Nicklin, C., & Plonsky, L. (2020). Outliers in L2 research in applied linguistics: A synthesis and data re-analysis. *Annual Review of Applied Linguistics*, *40*, 26–55. https://doi.org/10.1017/S0267190520000057

Pickering, M. J., & Majid, A. (2007). What are implicit causality and consequentiality? *Language and Cognitive Processes*, *22*, 780–788. https://doi.org/10.1080/01690960601119876

Pyykkönen, P., & Järvikivi, J. (2009). Activation and persistence of implicit causality information in spoken language comprehension. *Experimental Psychology*, *57*, 5–16. https://doi.org/10.1027/1618-3169/a000002

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Rigalleau, F., Caplan, D., & Baudiffier, V. (2004). New arguments in favour of an automatic gender pronominal process. *The Quarterly Journal of Experimental Psychology*, *57*, 893–933. https://doi.org/10.1080/02724980343000549

Roberts, L., Gullberg, M., & Indefrey, P. (2008). Online pronoun resolution in L2 discourse. L1 Influence and general learner effects. *Studies in Second Language Acquisition*, *30*, 333–357. https://doi.org/10.1017/S0272263108080480

Sorace, A. (2011). Pinning down the concept of “interface” in bilingualism. *Linguistic Approaches to Bilingualism*, *1*, 1–33. https://doi.org/10.1075/lab.1.1.01sor

Sorace, A., & Filiaci, F. (2006). Anaphora resolution in near-native speakers of Italian. *Second Language Research*, *22*, 339–368. https://doi.org/10.1191/0267658306sr271oa

Stanovich, K. E., & West, R. F. (1989). Exposure to print and orthographic processing. *Reading Research Quarterly*, *24*, 402–433. https://doi.org/10.2307/747605

https://doi.org/10.1017/S0272263122000468 Published online by Cambridge University Press

Van Berkum, J. J., De Goede, D., Van Alphen, P., Mulder, E., & Kerstholt, J. H. (2013). How robust is the language architecture? The case of mood. *Frontiers in Psychology*, *4*, 505. https://doi.org/10.3389/fpsyg.2013.00505

Van Dyke, J. A., & Johns, C. L. (2012). Memory interference as a determinant of language comprehension. *Language and Linguistics Compass*, *6*, 193–211. https://doi.org/10.1002/lnc3.330

Van Dyke, J. A., Johns, C. L., & Kukona, A. (2014). Low working memory capacity is only spuriously related to poor reading comprehension. *Cognition*, *131*, 373–403. https://doi.org/10.1016/j.cognition.2014.01.007

Van Dyke, J. A., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*, 157–166. https://doi.org/10.1016/j.jml.2006.03.007

Cite this article: Wang, T. and Gabriele, A. (2023). Individual differences modulate sensitivity to implicit causality bias in both native and nonnative processing. *Studies in Second Language Acquisition,* 45: 853–881. https://doi.org/10.1017/S0272263122000468