Journal of Memory and Language 127 (2022) 104368

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Still no evidence for audience design in syntax: Resumptive pronouns are not the exception☆

Adam M. Morgan [a,*], Victor S. Ferreira [b]

[a] NYU School of Medicine, Department of Neurology, 227 E 30th St, New York, NY 10016, USA
[b] UC San Diego, Department of Psychology, 9500 Gilman Dr., La Jolla, CA 92093, USA

ARTICLE INFO

Keywords:
Language production
Audience design
Resumptive pronouns

ABSTRACT

Speakers often tailor their speech to the needs of their interlocutors to facilitate comprehension. While such *audience design* can be observed in the words speakers choose (e.g., using simpler words when talking to children) or the volume of their voice (louder in noisier environments), it has rarely been observed in what syntactic structures speakers use. Consequently, many researchers have concluded that syntactic audience design is impossible. However, there exists a parallel literature in which syntactic audience design is assumed to play a central role. Specifically, English *resumptive pronouns* (e.g., …*pronouns that nobody knows why people say* ***them***) are puzzling in that, despite being ungrammatical, they are regularly and reliably produced. To account for this, some theories hold that speakers produce resumptive pronouns to improve the acceptability of their utterances. If so, this would be an exceptional case of syntactic audience design. In three experiments, we test this hypothesis by eliciting sentences from speakers while manipulating acceptability. We consistently show no effect of acceptability on rates of resumptive pronoun production, despite successfully manipulating production rates with a comparably sized control manipulation. We conclude that resumptive pronouns are not in fact the result of audience design, and therefore do not constitute a challenge to the idea that syntactic audience design is impossible.

Introduction

When talking to children, adult speakers tend to use shorter sentences, more common words, and exaggerated pitch contours (Snow, 1972). Speakers in noisy environments will involuntarily speak louder (Lombard, 1911) (a fact which early-20th Century physician Étienne Lombard famously tried to leverage to root out patients faking deafness to avoid work; Zollinger & Brumm, 2011). And when speakers have a choice between multiple words for the same object – *shoe* or *sneaker* – they tend to produce whichever word they previously used with the particular person they are speaking to, even if they used a different word more recently with a different interlocutor (Brennan & Clark, 1996).

All of these are examples of *audience design*, or tailoring speech for the benefit of the comprehender. Audience design in general is a necessary feature of a system like language, the goal of which is to ensure that the comprehender understands the message. Even so, not every decision that a speaker makes at each level in the system serves audience design, even when it may seem that speakers construct their utterances in particular ways so as to be better understood.

For example, in the domain of syntax, there is little evidence that speakers choose syntactic structures that benefit the comprehender. One common paradigm that has been used to probe for syntactic audience design looks at whether speakers avoid syntactic structures that result in ambiguous sentences. For instance, in "garden path sentences" like (1), the word *mayor* may initially be interpreted as the direct object of *believe*, as if the sentence were *The protestors believed the mayor*.

(1) The protestors believed the mayor would resign.

However, the next word, *would*, does not fit this parse. Upon encountering *would*, comprehenders who initially analyze *mayor* as a direct object must reanalyze it as the subject of *resign*, causing a

☆ This research was supported in part by the Graduate Research Fellowship Program at the US National Science Foundation under grant DGE-1144086 to Adam M. Morgan and by the National Institute of Child Health and Human Development under the US National Institutes of Health under grant R01 HD051030 to Victor S. Ferreira. Stimuli, data, analyses can be found online at <<https://osf.io/2aeuf/>>.

* Corresponding author.
E-mail address: adam.morgan@nyulangone.org (A.M. Morgan).

https://doi.org/10.1016/j.jml.2022.104368
Received 12 January 2022; Received in revised form 2 August 2022; Accepted 5 August 2022

temporary disruption in comprehension processes (Trueswell, Tanenhaus, & Kello, 1993).

If speakers choose particular syntactic structures according to the needs of their listeners, they should avoid garden path sentences like (1). For instance, the ambiguity could be avoided with the use of the optional complementizer *that*, as in *The protestors believe* **that** *the mayor would resign*. However, studies consistently find no evidence that speakers engage in this type of syntactic ambiguity avoidance (Arnold, Wasow, Asudeh, & Alrenga, 2004; Ferreira & Dell, 2000; Jaeger, 2010; Ferreira & Schotter, 2013; but see Elsness, 1984; Temperley et al., 2003).

In fact, there is sufficiently little evidence for syntactic audience design (or, for that matter, any type of executive control over syntactic processing; see Ferreira & Clifton Jr, 1986) that production models involving executive control and syntax do not posit an interface between the two, thereby rendering it impossible for speakers to directly choose a particular syntactic structure for the benefit of the comprehender (Levelt, 1993; Bock & Levelt, 1994; Ferreira, Morgan, & Slevc, 2018; Garrett, 1975). For instance, Bock's (1982) model includes an interface between executive control and phonetic encoding, reflecting speakers' ability to tailor acoustic properties to the needs of the listener (as in speaking louder in noisy environments). By excluding such a link between executive control and syntactic encoding, the model does not allow for the possibility of speakers modulating syntax to meet the listener's needs. More recent work (Ferreira, 2019; Jaeger & Ferreira, 2013; Kurumada & Jaeger, 2015; Buz, Tanenhaus, & Jaeger, 2016) has expanded upon these models in order to account for longer-term syntactic learning of the type suspected to underlie phenomena like syntactic priming (Pickering & Garrod, 2004; Ferreira, 2019). But even in the most complex models, there remains no direct mechanism for syntactic audience design.

A contradicting literature: Resumptive pronouns

While the production literature has generally converged on the idea that speakers do not engage in syntactic audience design, there exists a parallel literature in which it has largely been assumed that audience design accounts for an otherwise puzzling set of data. Specifically, recent work on a quixotic grammatical phenomenon in English known as *resumption* suggests that researchers may simply have been looking in the wrong place for evidence of syntactic audience design. Resumption, or the use of *resumptive pronouns* like *she* in (2), is often argued to be used in production for the benefit of the comprehender.

(2) That's the professor who I couldn't understand a word ***she*** said.

While native English speakers report that resumptive pronouns are unacceptable, they readily produce them. Indeed, the alternative – leaving out the pronoun and using a *gap* (indicated with an underscore throughout) – is often even less acceptable, as in (3).

(3) That's the professor who I couldn't understand a word _ said.

Resumptive pronouns have long been of interest to linguists and psycholinguists because they challenge traditional notions of grammaticality. That is, there are two standard metrics for whether or not a structure is grammatical: production and comprehension. If native speakers regularly produce a particular structure, it is generally assumed to be grammatical. Similarly, if native speakers deem a structure to be highly acceptable, then it is also generally assumed to be grammatical.

In many unrelated languages, this is how resumption works. Speakers of Lebanese Arabic (Aoun, Choueiri, & Hornstein, 2001), Chinese (Pan, 2016), Gbadi (Koopman, 1983), Hebrew (Shlonsky, 1992), Irish (McCloskey, 1990), and Swedish (Engdahl, 1985) find resumptive pronouns highly acceptable, and also readily produce them in speech. However, the pattern of data in English breaks the mold. In spite of finding resumptive pronouns highly unacceptable, English speakers produce them quite frequently (Bennett, 2009). At least four experimental studies have demonstrated that they can be reliably elicited in the lab (Ferreira & Swets, 2005; Morgan & Wagers, 2018; Fadlon, Morgan, Meltzer-Asscher, & Ferreira, 2019; Morgan, von der Malsburg, Ferreira, & Wittenberg, 2020).

To account for this apparent paradox, previous researchers have appealed to notions of audience design. The idea is pervasive in the theoretical literature (Kroch, 1981; Ackerman, Frazier, & Yoshida, 2018; cf. Keffala & Goodall, 2011; Han et al., 2012), but was formalized by Morgan and Wagers (2018) to account for their finding that speakers' propensity to produce a resumptive pronoun was directly related to acceptability. Specifically, they proposed that speakers produce resumptive pronouns to improve the acceptability of their utterances for the benefit of the comprehender. The aim of the present paper is to test this hypothesis directly, by manipulating acceptability and looking for corresponding changes in how many resumptive pronouns speakers produce.

To understand Morgan and Wagers's (2018) study, a bit of background on the syntax of resumption is necessary. Resumptive pronouns occur in *long-distance dependencies*, where a noun appears far away from its canonical position. For instance, in ordinary English sentences like (4a), the subject (*dog*) generally closely precedes the verb (*ate*), and the object (*couch*) closely follows it. (This structure will be referred to as a *non-island*, in contrast to *island* structures, which are explained below.) In long-distance dependencies, like (4b) and (4c), nouns can appear far away from their canonical position. In (4b), *dog* appears two clauses (indicated with brackets) away from its canonical position. Similarly, in (4c), *couch* appears two clauses away from its canonical position.

(4) *Non-island*
   a. The owner thinks [the dog ate the couch].
   b. That's the **dog** [that the owner thinks [_ ate the couch]].
   c. That's the **couch** [that the owner thinks [the dog ate _]].

Long-distance dependencies are implicated in a variety of structures, including *wh-questions* (e.g., *What did the owner think the dog ate _?*), *tough-constructions* (*The couch was tough to chew _*), topicalization (*I'm not a big coffee person, but tea I can't stand _*), and others. Throughout this paper, we will use relative clauses to create long-distance dependencies, as in (4b) and (4c). Relative clauses are clauses that modify a noun. The modified noun, or *head noun* – *dog* in (4b) and *couch* in (4c) – is not repeated inside the relative clause, leaving a gap.

English resumptive pronouns are extremely uncommon in non-islands like (4). Speakers nearly always leave gaps in these structures, as in (4b) and (4c). However, there exists a class of structures where resumptive pronouns are more common: *islands*. On their own, islands structures are perfectly grammatical, as in (5a), an example of a type of island called a *weak* island. Islands' defining feature is that when gaps appear inside of them, they become unacceptable. Thus, while one can create an acceptable relative clause out of a non-island, as in (4b) and (4c), one cannot do the same with island structures, as demonstrated in (5b) and (5c).

(5) *Weak island*
   a. The owner wondered when the dog ate the couch.
   b. *That's the dog that the owner wondered when _ ate the couch.
   c. ?That's the couch that the owner wondered when the dog ate _.

Exactly how unacceptable these *island violations* are depends on two properties of the sentence. The first is the syntactic position of the gap: gaps in subject position, or *subject gaps*, are less acceptable than *object gaps* (but see Morgan, 2022 for a more nuanced view). Thus, (5b) is less acceptable than its counterpart, (5c). This graded acceptability is indicated throughout with question marks and asterisks: no marking

indicates that a sentence is fully acceptable, and question marks, single asterisks, and double asterisks indicate increasing degrees of unacceptability.

The second relevant property is the particular type of island structure. Islands include a conglomeration of different syntactic structures, all of which become unacceptable with a gap inside of them. These structures naturally cluster into two subcategories: First, *weak islands* (e.g., 5), which are defined as including structures like embedded interrogative clauses and complex noun phrases, are moderately unacceptable with gaps in them. Second, *strong islands* (e.g., 6), which are defined as including structures like adjoined clauses and relative clauses, are severely unacceptable with gaps in them (e.g., Han et al., 2012; Keffala & Goodall, 2011; Morgan & Wagers, 2018). Note that while *weak* and *strong* are labels for groups of syntactic structures with similar acceptability profiles, throughout this paper we use only two types of island structures: embedded interrogative clauses, which we will refer to as "weak islands," and adjunct clauses, which we will refer to as "strong islands."[1]

(6) *Strong island*
   a. The owner cried because the dog ate the couch.
   b. **That's the dog [that the owner cried because [_ ate the couch]].
   c. *That's the couch [that the owner cried because [the dog ate _]].

Based on the fact that resumptive pronouns primarily appear in island structures, it is tempting to suggest that speakers use them to ameliorate sentences that would be even worse if they contained a gap. What exactly resumptive pronouns might ameliorate remains a source of disagreement. Early proposals mainly focused on acceptability – the hypothesis was that (6b) might sound better if it had a resumptive pronoun in it, as in *That's the dog that the owner cried because* **it** *ate the couch.* However, numerous acceptability judgment studies have found that resumptive pronouns very rarely render a sentence more acceptable than the corresponding sentence with a gap (Han et al., 2012; Keffala & Goodall, 2011; Heestand, Xiang, & Polinsky, 2011), even in sentences where speakers commonly produce them (Morgan & Wagers, 2018) (although see Ackerman et al., 2018 for conflicting evidence from a forced-choice paradigm).

Other proposals hold that resumptive pronouns do not serve to improve acceptability per se, but to facilitate processing (Prince, 1990). Some argue that the benefit is for the speaker: the use of a pronoun may somehow reduce the burden on the production system (Polinsky, Clemens, Morgan, Xiang, & Heestand, 2013; Fadlon et al., 2019; Asudeh, 2011; Asudeh, 2012). Most, however, hold that speakers produce resumptive pronouns not for their own benefit, but to help listeners glean the intended meaning (Hofmeister & Norcliffe, 2013; Beltrama & Xiang, 2016; Hammerly, 2019; Zenker & Schwartz, 2021). For instance, several studies have reported evidence that sentences with resumptive pronouns are read faster than sentences with gaps, hinting at a potential facilitatory effect in comprehension (Hofmeister & Norcliffe, 2013; Hammerly, 2019; Zenker & Schwartz, 2021).

However, there are some reasons to doubt these findings. For one, the facilitatory effect is usually quantified as an increase in reading speed – e.g., words per second. But given that sentences with resumptive pronouns have one more word than sentences with gaps, even though readers speed up, they still take longer overall to process the same semantic content. Furthermore, even in sentences where speakers produce resumption, resumptive pronouns in fact seem to *hinder* comprehension, not help (Morgan et al., 2020; although see Zenker & Schwartz, 2021).

Much of the debate hinges on the distribution of resumptive pronouns in production. For instance, if resumptive pronouns serve to ameliorate acceptability, one should expect speakers to produce them in precisely those circumstances in which a resumptive pronoun would render a sentence more acceptable than a gap. To test this, Morgan and Wagers (2018) ran two experiments. In an acceptability judgment experiment, they asked participants to rate the acceptability of various types of sentences with gaps or resumptive pronouns. In a production experiment, they elicited the same sentences from speakers to see how commonly resumptive pronouns were produced in each sentence type. The sentences were manipulated in two ways. First, they had various structures – specifically, two types of non-island structures and four types of island structures. They also manipulated the syntactic position of the gap – whether it appeared in subject position, object position, or in a third, "matrix" position (e.g., *That's the owner who _ wondered when the dog ate the couch*). In all, they tested 15 distinct sentence types that spanned a wide range of acceptabilities.

They found that the rate of resumptive pronoun production in a given structure was predicted by two factors: *gap acceptability*, or the acceptability of that structure with a gap in it, and whether or not that structure was an island. Specifically, the lower the acceptability of the gap structure, the more likely participants were to produce a resumptive pronoun. Speakers were even more likely to use a resumptive pronoun if the structure was an island. Interestingly, *resumptive pronoun acceptability*, or the acceptability of a structure with a resumptive pronoun in it, did not significantly predict production.

To account for these findings, Morgan and Wagers, 2018 adapted a model by Asudeh (2012), according to which resumptive pronouns are the result of production processes breaking down. They propose that resumption has two triggers: acceptability and grammaticality. Specifically, to produce a long-distance dependency, the production system must keep track of the head noun until at least its canonical interpretation site so it knows where to leave behind a gap (Fadlon et al., 2019). Prior to articulation, the production system plans the dependency with a gap – the default strategy for long-distance dependencies. The system then assesses the acceptability of that planned sentence. The more unacceptable that planned utterance is, the more likely the system is to give up on maintaining the dependency. When the system does give up, it defaults to producing all nouns overtly. This results in a resumptive pronoun. To account for the increased rates of resumption in islands relative to non-islands (above what acceptability alone predicts), they suggest that ungrammaticality can also trigger this breakdown.

This model accounts for the correlation between gap acceptability and resumptive pronoun production by assuming that acceptability plays a causal role in production. It would also explain why resumptive pronoun acceptability does not correlate with production: gaps are the default (and perhaps only) grammatical strategy for forming long-distance dependencies in English. Resumptive pronouns are just what happens when the system breaks down, giving up on a proper long-distance dependency.

What is important to note about this model is that the reliance on acceptability implies a degree of audience design. That is, acceptability serves no purpose in production, but may be an important feature to monitor for the sake of the interlocutor, for whom it may index comprehensibility (Beltrama & Xiang, 2016), or even the interlocutors social perceptions of the speaker.

The reliance on audience design is not new in the resumption literature. Indeed, nearly all previous accounts assume some version of audience design (see, e.g., Beltrama & Xiang, 2016; Dickey, 1996; Hofmeister & Norcliffe, 2013; Prince, 1990), and it is easy to imagine why. Relative to gaps, pronouns can provide overt person, number, and animacy features that might help the listener track the intended meaning. This could be especially helpful in processing complex sentences like islands, where it might be harder to keep track of the dependency.

[1] These particular structures – embedded interrogatives and adjuncts – were chosen for two reasons. First, they are commonly used in the experimental literature (see, e.g., Morgan et al., 2020), facilitating comparison of our study to previous research. Second, they span a broad range of acceptability, thus boosting the internal validity of our experiments.

But, as outlined above, as standardly formulated, production models should make producing a resumptive pronoun for the sake of the interlocutor impossible. If these production models are correct, a theory of resumption is in order which does not rely on audience design (see, e.g., Asudeh, 2012; Polinsky et al., 2013; Morgan et al., 2020). On the other hand, if resumption models with audience design as a central component are correct, it indicates a need to update the basic architecture of standard production models.

The present work: Testing the Acceptability Hypothesis

This paper therefore directly tests the hypothesis that speakers produce resumptive pronouns to improve the acceptability of a sentence – what we call the *Acceptability Hypothesis* of resumption – by taking a finer-grained look at the relationship between acceptability and resumptive pronoun production. Specifically, Morgan and Wagers, 2018 manipulated acceptability by changing the syntactic structure of their sentences, using non-islands and islands like (4), (5) and (6). Consequently, acceptability and syntactic structure were entirely confounded.

This leaves open a number of other possible explanations of resumptive pronoun production, including virtually any other property of structures that correlates with acceptability. Reasonable candidates might be features like structural frequency, complexity, working memory burden, or processing difficulty (Hofmeister, Jaeger, Arnon, Sag, & Snider, 2013; Staum Casasanto, Hofmeister, & Sag, 2010; Hofmeister, Casasanto, & Sag, 2014). Whereas it is hard to imagine a reason that acceptability might play a causal role other than audience design, constructs like these can clearly impact production processes for reasons that have nothing to do with the interlocutor. We will return to these possibilities in the General Discussion.

Below we present three experiments that test the key prediction of the Acceptability Hypothesis: that changes to gap acceptability should lead to corresponding changes in resumptive pronoun production, even when holding the global syntactic structure of the sentence constant. If, on the other hand, a feature like structural frequency is the underlying cause of resumptive pronoun production, then changes to acceptability that do not impact syntax should not impact rates of resumption.

To test this, we deconfounded syntax and acceptability by leveraging the fact that the particular words in a sentence, or the sentence's *lexicalization*, can impact acceptability independent of the global syntactic structure. For instance, Bever (1970, 1974) demonstrated that if the nouns in a complex sentence are more similar to one another, the sentence is less acceptable. Thus, while (7a) and (7b) have the same global syntactic structure, (7a), which has mutually distinctive nouns (one full, referential determiner phrase, *the dog*; one quantified nominal, *someone*; and one local-person pronoun, *I*), is more acceptable than (7b), in which the nouns are mutually similar (they are all full, referential determiner phrases).

(7) a.
The dog that someone I know adopted ran away.
b. The dog that the senator the teacher knows adopted ran away.

Experiment 1, an acceptability judgment study, confirms that syntactic and lexical similarity manipulations (like those in Morgan & Wagers (2018) and Bever (1974), respectively) independently impact acceptability. Exps. 2 and 3 are production experiments which elicit long-distance dependencies from participants using the manipulations validated in Exp. 1.

Experiment 1: Acceptability judgment

Prior to testing whether unacceptability causes resumptive pronoun production, we performed an acceptability judgment study to validate our stimuli, and in particular to verify that our acceptability manipulations affect acceptability in the expected way. We aimed to replicate two findings: first, that mutual lexical similarity decreases acceptability relative to mutual distinctiveness (as in Bever, 1974), and (2) that syntactic structure impacts acceptability – specifically, that non-islands are more acceptable than weak islands, which are in turn more acceptable than strong islands (as in Morgan & Wagers, 2018).

Table 1
Sample stimulus item from Exp. 1.

SYNTAX	LEXICALIZATION	Stimulus
Non-island	*Distinct*	It was the little boy who we thought that Willy Wonka heard that someone hit _.
	Similar	It was the little boy who the clown thought that the clerk heard that the oompa loompa hit _.
Weak island	*Distinct*	It was the little boy who we wondered why someone hit _.
	Similar	It was the little boy who the clown wondered why the oompa loompa hit _.
Strong isl.	*Distinct*	It was the little boy who we screamed when someone hit _.
	Similar	It was the little boy who the clown screamed when the oompa loompa hit _.

Data availability

For all three experiments, stimuli, data, analyses can be found online at <<https://osf.io/2aeuf/>>.

Method

Participants

35 participants were recruited from Amazon's Mechanical Turk. Data from one self-identified native bilingual participant were discarded prior to analysis. Of the 34 participants included in the analysis, mean age was 35.5 (s.d. = 11), and all were native monolingual speakers of American English.

Design

Two factors were manipulated, resulting in a 2×3 design (see Table 1 for a sample item set). The first, LEXICALIZATION, had two levels: *similar* and *distinct*. This was achieved as in (7) following Bever (1974), with stimuli in the *similar* condition containing only full, referential determiner phrases (e.g., *the dog*), and stimuli in the *distinct* condition each containing at most one of four different types of nominals, including full determiner phrases, local person pronouns (e.g., *you*),[2] quantified nominals (e.g., *someone*), and names (e.g., *Becca*). Following Bever (1974) and Lewis (1996), we expected *distinct* sentences to be rated more acceptable than *similar* sentences – a main effect of LEXICALIZATION.

The second factor, SYNTAX, had three levels: *non-island*, *weak island*, and *strong island*. The two island conditions had exactly the same structures as Examples (5) and (6). The *non-island* condition was similar to Ex. (4), but it contained an additional level of clausal embedding (i.e., rather than just *'…the little boy [who Willy Wonka heard [that someone hit]]'*, an extra clause boundary interceded between the head noun and gap: *'…the little boy [who we thought [that Willy Wonka heard [that someone hit]]]'*). This feature was included to boost the baseline rate of

[2] Reviewers pointed out that the use of local person pronouns such as *I* and *you* in our *distinct* conditions introduces an additional distinction in information structure, which previous work has shown affects how long-distance dependencies are processed (Warren & Gibson, 2005). In the appendix we report supplementary analyses which show that excluding items with local person pronouns does not change the pattern of results. This indicates that whatever additional differences are introduced by the use of local pronouns, these do not impact our conclusions.

resumptive pronoun production in non-islands. Morgan and Wagers (2018) found almost no resumptive pronoun production in non-islands with just one level of embedding, which would make any effect of LEXICALIZATION in non-islands difficult to observe. By adding a second level of embedding to their non-island stimuli, they increased rates of resumption in production. We aimed to emulate that here.

Previous literature indicates that structures fall on an acceptability cline, starting with non-islands (most acceptable), then weak islands, then strong islands (Dickey, 1996; McDaniel & Cowart, 1999; Alexopoulou & Keller, 2007; Heestand et al., 2011; Keffala & Goodall, 2011; Han et al., 2012; Clemens, Morgan, Polinsky, & Xiang, 2012; Polinsky et al., 2013; Morgan & Wagers, 2018), at least when these structures all have the same number of clauses. We predict that in our stimuli, strong islands will be less acceptable than weak islands. However, it is less clear what to expect for our non-island stimuli. This is because of the extra level of embedding in the non-islands, which severely reduces acceptability. Indeed, Morgan and Wagers (2018) found that in terms of acceptability, doubly-embedded non-islands were comparable to singly-embedded weak islands. Given that doubly-embedded structures have not been extensively studied in the literature,[3] we make no predictions for the acceptability of our *non-island* stimuli. It is important to note that whatever the outcome, is inconsequential for our main question – whether the resumptive pronoun production is caused by acceptability. If the LEXICALIZATION manipulation affects the acceptability of the non-island stimuli in the expected way, then we will still be able to address this question.[4]

Each participant saw 72 fillers and 36 critical items in a pseudo-random order, with conditions counterbalanced across item sets and lists created according to a Latin-square design. A full set of conditions for one experimental item is given in Table 1; the rest (along with all data and analyses) may be found on the OSF repository at https://osf.io/2aeuf/.

Procedure

The experiment was run online in Ibex Farm (Drummond, 2013) and took roughly 25 min to complete. Participants were instructed to rate the stimuli on a 1–9 Likert scale and not to worry "about the kinds of grammar rules you may have learned in high school English classes." They were instead encouraged to try to base "ratings on factors like how natural a sentence sounds, or whether you might expect a native English speaker to say it given an appropriate context."

Analysis

Prior to analysis, trials in which the sentence was read and rated in less than 2046 ms (the bottom 2% of reaction times) were excluded to remove data where participants probably did not carefully read the sentence. Raw ratings were *z*-scored to remove extraneous differences arising from, for example, individual differences in use of the Likert scale. Using the *lme4* package in R (Bates, Mächler, Bolker, & Walker, 2014; R Core Team, 2021), a linear mixed effects regression was fit to the data with fixed effects for LEXICALIZATION and SYNTAX and their interaction. In this and all subsequent models, fixed effects were treatment coded and subjects and items were included as random intercepts, with every fixed effect allowed to vary with every random intercept. If the model did not converge, random correlations were removed and then random slopes, one by one in order of least variance to most, until the model did converge (Barr, Levy, Scheepers, & Tily, 2013). The final random effects structures for each model in the paper are reported with the results. Significance was assessed using the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017).

[3] The only other study we are aware of that uses such doubly-embedded structures is Morgan (2022), and this work also found that they are much less acceptable than their singly-embedded counterparts.

[4] A third manipulation was also included in Exp. 1, but is not discussed here for the sake of simplicity as it was deemed not to contribute anything substantive. Specifically, DEPENDENCY type – whether the long-distance dependency was realized with a *gap* or a resumptive pronoun (*RP*) – was manipulated in order to fully replicate the design of Morgan and Wagers (2018). However, as Morgan and Wagers (2018) demonstrate with a bootstrapping analysis, the acceptability of the sentence with resumptive pronouns in it does not predict resumptive pronoun production; only the acceptability of the sentence with a gap in it. Therefore, we deemed this condition unnecessary to addressing our critical question and removed *RP* conditions from all analyses and discussion below. The full dataset, including *RP* conditions, is available on OSF at https://osf.io/2aeuf/.

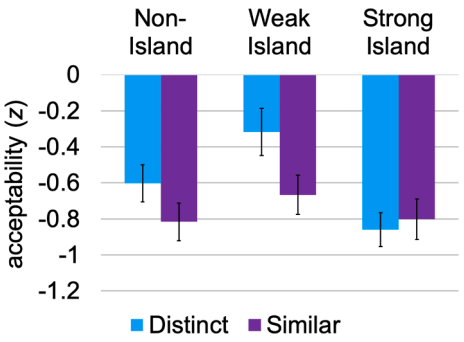

Fig. 1. Mean acceptability rating (*z*-scored) and standard error, by condition.

Table 2
Experiment 1: Mean *z*-scored acceptability ratings (and standard error) by condition. Effect sizes for LEXICALIZATION are given for each syntactic structure in the bottom row of the table. The effect sizes (Cohen's *d*) for SYNTAX (collapsing across LEXICALIZATION) were: 0.327 for non- vs. weak islands; 0.198 for non- vs. strong islands; and 0.508 for weak vs. strong islands.

	non-island		*weak island*		*strong island*	
distinct	−0.603	(0.102)	−0.318	(0.131)	−0.860	(0.094)
similar	−0.817	(0.103)	−0.666	(0.109)	−0.801	(0.113)
Cohen's *d*	0.305		0.581		0.096[†]	

[†] Note that this effect size reflects a difference which we argue is confounded by a floor effect, and as such should not be interpreted directly.

Table 3
Experiment 1: Statistical results.

	β	d.f.	*t*	*p*	
Intercept (*weak island, distinct*)	−0.327	147.445	−4.756	< .001	***
SYNTAX (*non-island, distinct*)	−0.277	143.497	−3.070	.003	**
SYNTAX (*strong island, distinct*)	−0.512	146.342	−5.695	< .001	***
LEXICALIZATION (*weak, similar*)	−0.328	223.269	−3.882	< .001	***
Interaction (*non-island, similar*)	0.140	448.400	1.170	.243	
Interaction (*strong isl., similar*)	0.346	453.604	2.895	.004	**

Results

Mean *z*-scores of ratings for each condition are shown in Fig. 1 and given in Table 2. Ratings for critical stimuli were low overall (median *z* was −0.76), which was expected given that the stimuli consisted of island sentences, which are ungrammatical, and non-island sentences with two clausal embeddings, which, while technically grammatical, are highly degraded. (By comparison, filler sentences were matched for length, but were all grammatical and less complex.).

Results of the model, which converged with the full random effects structure, are shown in Table 3. As expected, strong islands were rated significantly worse than weak islands ($\beta = -0.512, t = -5.695, p < .001$). Interestingly, non-islands were also rated significantly worse than weak islands ($\beta = -0.277, t = -3.070, p = .003$). Critically, as predicted, there was a significant effect of LEXICALIZATION: sentences with

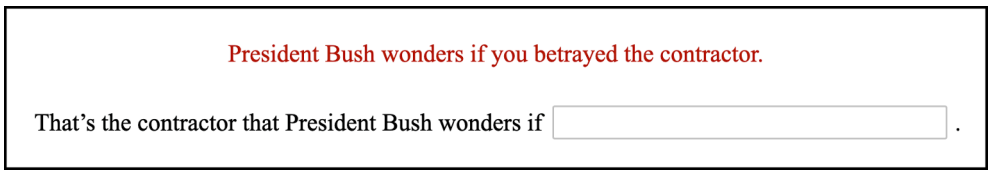

Fig. 2. Screenshot from Exp. 2. The upper sentence is the BASE, the lower one the PREAMBLE. Participants were instructed to complete the lower sentence using the information given in the upper one.

mutually similar nominals were rated less acceptable than sentences with mutually distinctive nominals ($\beta = -0.328$, $t = -3.882$, $p < .001$). Finally, there was a significant interaction term reflecting the fact that in strong islands, lexical similarity did not decrease acceptability ($\beta = 0.346, t = 2.895, p = .004$).

Discussion

Results were largely as predicted. Interestingly, non-islands were rated significantly *worse* than weak islands. As mentioned above, this likely reflects the extra level of clausal embedding in the non-islands relative to the other two conditions, but does not bear on our main question.

Crucially, the LEXICALIZATION manipulation had the predicted effect on acceptability: increased lexical similarity decreased acceptability, at least in in *non-island* and *weak island* conditions. In the *strong island* condition, however, this effect disappeared. We believe that this most likely reflects a floor effect – that is, the 1–9 point Likert scale may not have been sufficient to capture underlying differences between the least acceptable conditions. Strong islands have extremely low acceptability (indeed the modal rating in these conditions was 1), and participants may not have had enough room at the bottom of the scale to distinguish between gradations of *extremely low*. However, even if the data are to be taken at face-value and the LEXICALIZATION manipulation does not impact acceptability in strong islands, the finding of an effect of LEXICALIZATION in non-islands and weak islands means that our main hypothesis can be tested by looking at how the LEXICALIZATION manipulation affects production within these two structures.

Specifically, if resumptive pronoun production is indeed the result of low acceptability, then we should expect to see higher rates of resumptive pronoun production in *similar* conditions than in *distinct* conditions, at least in non-islands and weak islands. Experiments 2 and 3 were designed to test this prediction.

Experiment 2: Written Production

Experiment 2 was a written production task that elicited sentences with gaps or resumptive pronouns. The goal was to test whether speakers produce resumptive pronouns to improve the acceptability of their utterances – i.e., the Acceptability Hypothesis. If so, speakers should produce more resumptive pronouns in sentences with higher lexical similarity (which have lower acceptability) than in sentences with more distinct words (which have higher acceptability).

Method

Participants

33 participants were recruited via Amazon's Mechanical Turk. Data from five self-identified native bilingual participants were excluded prior to analysis. Of the 28 participants included in the analysis, mean age was 36 (s.d. = 10) and all were native monolingual speakers of American English.

Design

The study design was modeled after Morgan and Wagers (2018), using the same paradigm and stimuli derived from theirs. In each trial, participants were shown two lines of text and a text box, as shown in Fig. 2. The first line of text was a *base sentence* (in Fig. 2: *President Bush wonders if you betrayed the contractor*). The base sentence did not involve a relative clause or other long-distance dependency. All nouns therefore appeared in their canonical positions.

The second line, the PREAMBLE, introduced a relative clause (in Fig. 2: *That's the contractor that…*). The head noun (*contractor*) was always identical to the lowest object in the base sentence. The starter sentence ended at the lowest complementizer. Participants were instructed to complete the second sentence as a paraphrase of the first, thereby requiring them to produce a gap or resumptive pronoun.

The stimuli, demonstrated in Table 4, appeared in the same 3 × 2

Table 4
Sample stimulus item from Exp. 2. For each cell, the base sentence appears in the top row, then the preamble (in italics), and then the expected response (in italics and parentheses).

Non-island	
	Somebody thought that you implied that Dr. House harassed the orderly.
Distinct	*I sided with the orderly who somebody thought that you implied that…*
	(…Dr. House harassed _/him.)
	The doctor thought that the judge implied that the nurse harassed the orderly.
Similar	*I sided with the orderly who the doctor thought that the judge implied that…*
	(…the nurse harassed _/him.)
Weak island	
	Somebody learned when Dr. House harassed the orderly.
Distinct	*I sided with the orderly who somebody learned when…*
	(…Dr. House harassed _/him.)
	The doctor learned when the nurse harassed the orderly.
Similar	*I sided with the orderly who the doctor learned when…*
	(…the nurse harassed _/him.)
Strong island	
	Somebody filed a complaint when Dr. House harassed the orderly.
Distinct	*I sided with the orderly who somebody filed a complaint when…*
	(…Dr. House harassed _/him.)
	The doctor filed a complaint when the nurse harassed the orderly.
Similar	*I sided with the orderly who the doctor filed a complaint when…*
	(…the nurse harassed _/him.)

Table 5
Examples of the three types of responses excluded from Exp. 2 analysis.

Passive (72)	
BASE	We screamed when some guy hit the little boy.
PREAMBLE	It was the little boy who we screamed when…
RESPONSE	*was hit by some guy.*
Noun repetition (35)	
BASE	The reporter vomited while the alien dissected the lunatic.
PREAMBLE	I knew the lunatic who the reporter vomited while…
RESPONSE	*the alien dissected the lunatic.*
Anomalous (73)	
BASE	Everybody closed shop while Jesse James spied on the jeweler.
PREAMBLE	I met the jeweler who everybody closed shop while…
RESPONSE	*he was spied on Jesse James.*

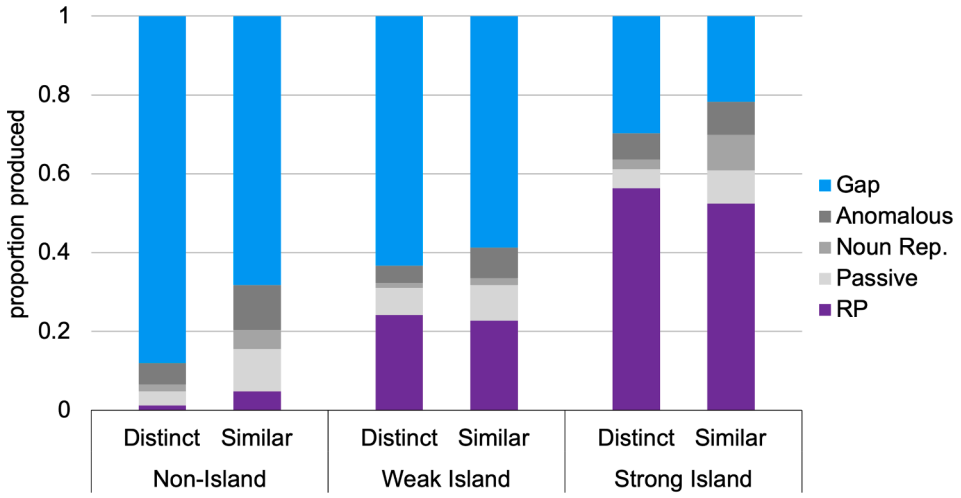

Fig. 3. Experiment 2: Proportions of each coded response type per experimental condition.

design as Exp. 1. The SYNTAX manipulation had three levels (*non-island*, *weak island*, and *strong island*), which were fully crossed with two levels of LEXICALIZATION (*similar* and *distinct*). Each participant saw 36 critical items and 60 fillers in a pseudo-random order.

Procedure

The experiment was programmed and run in Ibex Farm (Drummond, 2013) and took roughly one hour to complete. Prior to beginning 3 practice trials and then continuing on to the main experiment, participants were instructed that they would see two lines in each trial: a full sentence, in red font, above the beginning of a paraphrase of that sentence. They were told to complete the paraphrase by typing in the text box using information provided in the first sentence.

Data coding and analysis

Data from a total of 1,100 critical trials were collected.[5] Data from 107 trials were excluded because participants either accidentally skipped or made no effort to respond (e.g., "jkjk"). The remaining 993 experimental trials were coded for whether participants completed the relative clause with a gap (546 observations), resumptive pronoun (266 observations), or something else. A trial was coded as a gap or resumptive pronoun if the response matched the expected response (see Table 4), allowing for minor discrepancies like changes to the tense/aspect (e.g., *had attacked* rather than *attacked*), typos, perspective switching with local person pronouns (i.e., changing *I* to *you*), etc.

Changes that affected the event semantics or global syntactic structure were not coded as gap or resumptive pronoun, and were instead coded into three additional categories (see examples in Table 5): *passives* (72 observations), where the response clause was passivized so that the gap or resumptive pronoun appeared in subject position; *noun repetition* (35 observations), where rather than a gap or resumptive pronoun, the full noun was repeated in the canonical position; and *anomalous* responses (74 observations), which included semantically and/or syntactically incorrect responses, as well as any other responses that did not clearly fit into the previous categories.

In Experiments 2 and 3, all responses were coded independently by two undergraduate research assistants. The coders agreed on over 94% of trials in both experiments. Author AMM manually checked and coded trials on which the coders disagreed, and spot checked trials on which coders agreed.

[5] Due to a coding error, the first 15 participants only saw 35 of the 36 critical items. This was corrected mid-way through experiment running and the remaining 18 participants saw all 36 critical items.

All remaining analyses we report used logistic mixed effects regressions (Bates et al., 2014). Fixed effects, random effects, and our approach to achieving model convergence were all the same as in Exp. 1.

Results

The data are broken down by response type and condition in Fig. 3. Descriptively, there appears to be an effect of LEXICALIZATION in the rates of gap production. Interestingly, however, there does not appear to be a corresponding effect in rates of resumptive pronoun production. Instead, when lexical similarity was increased, rates of gap production seem to trade off with rates of production of passives, noun repetition, and anomalous responses (i.e., the responses that appear in shades of grey in Fig. 3; collectively referred to as 'other' responses throughout).

To test this characterization, we analyzed the data in a series of two nested models. First, we analyzed resumptive pronoun responses (coded as 1) vs. all other responses (coded as 0). This model converged with random intercepts for items and participants but no random slopes. Results, given in Table 6, show that participants produced more resumptive pronouns in weak islands than in non-islands ($\beta = -3.629$, $z = -4.799, p < .001$), and more resumptive pronouns in strong islands than weak islands ($\beta = 1.712, z = 5.977, p < .001$), consistent with the findings of Morgan and Wagers (2018). Critically, there was no effect of LEXICALIZATION on resumptive pronoun rate ($\beta = -0.064, z = -0.216, p =$

Table 6
Experiment 2 results. Model 1: The dependent variable was whether the response was a resumptive pronoun responses (coded as 1) vs. all other responses (coded as 0). Model 2: After excluding resumptive pronoun responses from the data, the dependent variable was whether the trial had a gap response (coded as 0) vs. 'other' response (i.e., passive, noun repetition, and anomalous responses; all coded as 1).

Model 1: RPs vs. gaps & 'other'	β	z	p	
Intercept (*weak island, distinct*)	−1.410	−4.712	< .001	***
SYNTAX (*non-island, distinct*)	−3.629	−4.799	< .001	***
SYNTAX (*strong island, distinct*)	1.712	5.977	< .001	***
LEXICALIZATION (*weak island, similar*)	−0.064	−0.216	.829	
Interaction (*non-island, similar*)	1.513	1.750	.080	.
Interaction (*strong island, similar*)	−0.172	−0.437	.662	
Model 2: 'Other' vs. gaps	β	z	p	
Intercept (*weak island, distinct*)	−2.378	−5.335	< .001	***
SYNTAX (*non-island, distinct*)	−0.450	−1.065	.287	
SYNTAX (*strong island, distinct*)	1.288	2.862	.004	**
LEXICALIZATION (*weak island, similar*)	0.683	1.746	.081	.
Interaction (*non-island, similar*)	0.743	1.390	.165	
Interaction (*strong island, similar*)	0.589	1.007	.314	

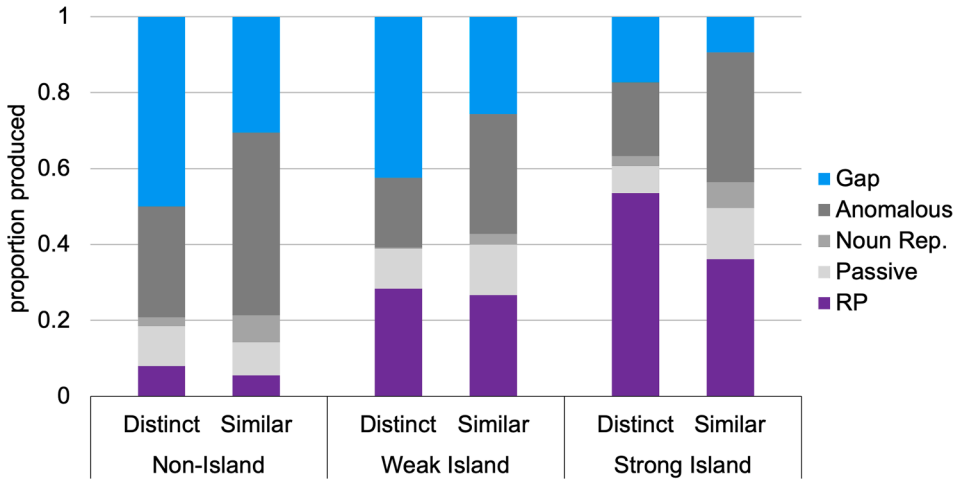

Fig. 4. Experiment 3: Proportions of each coded response type per experimental condition.

.829), although a marginal interaction reflected the fact that in non-islands, more resumptive pronouns were produced when nouns were mutually *similar* than when they were *distinct* ($\beta = 1.513, z = 1.750, p = .080$).

In Model 2, we statistically evaluated our observation that LEXICALIZATION appeared to have an effect on production, just not on the production of resumptive pronouns. Specifically, we looked for an effect of LEXICALIZATION on gap responses by excluding resumptive pronoun responses and comparing gap responses (coded as 0) to 'other' responses (including passive, noun repetition, and anomalous responses, all coded as 1). This model also converged with random intercepts for items and participants but no random slopes. The results revealed more 'other' responses in strong islands than in weak islands ($\beta = 1.288, z = 1.288, p = .004$). A marginal main effect of LEXICALIZATION reflected the fact that there were more 'other' responses in *similar* conditions than *distinct* conditions ($\beta = 0.683, z = 1.746, p = .081$), lending weak support to the observation that increased lexical similarity leads to a trade-off between gap and 'other' responses, but not resumption responses.

Discussion

Experiment 2 found that more resumptive pronouns are produced (a) in islands than in non-islands, and (b) in strong islands than in weak islands, replicating previous findings (Morgan & Wagers, 2018). Our critical question was whether increasing lexical similarity would result in more resumptive pronouns. No such effect was observed.

While we take caution in not directly interpreting statistically marginal results, two marginal findings are worth highlighting. One reflects the fact that in non-islands, there were numerically more resumptive pronouns produced when nouns were similar (4.8% resumption) than when nouns were distinct (1.2%). Given the marginal nature of this result, and the null findings in the weak island and strong island conditions, we think that this difference likely reflects noise in the data and not a true underlying effect. We can look at an independent replication of this comparison in Exp. 3.

Another interesting result is that while LEXICALIZATION did not affect rates of resumption (Model 1), Model 2 revealed a marginal effect reflecting fewer gaps and more 'other' responses in *similar* conditions relative to *distinct*.

Experiment 2 thus fails to find evidence for the Acceptability Hypothesis. However, as with any null result, the lack of evidence should not immediately be interpreted as evidence that there is no effect. In particular, Exp. 2 may have had insufficient power to detect a LEXICALIZATION effect. We argue, however, that there is reason to set aside such concerns. Specifically, in Exp. 1, the differences in acceptability induced by LEXICALIZATION (Cohen's *d* values of 0.305 and 0.581; see Table 2) were of comparable size to the differences in acceptability induced by the SYNTAX manipulation (Cohen's *d* values of 0.327, 0.198, and 0.508; see the caption of Table 2). If acceptability drives the different rates of resumptive pronoun production in different syntactic structures, then the highly significant SYNTAX effect in Exp. 2 indicates that we indeed had sufficient power to detect effects at this scale.

Nonetheless, to err on the side of caution, we aimed to address the potential lack of power issue with Exp. 3. One of the most powerful ways of arguing for the null hypothesis within the frequentist framework is with successful replication, which increases the odds that a null result reflects a true absence of an effect (Hoenig & Heisey, 2001). We therefore ran Exp. 3, a second production experiment, and boosted our experimental power by increasing the number of participants to 48.

Experiment 3 also aimed to address concerns about the ecological validity of Exp. 2. Specifically, while resumptive pronouns can be successfully elicited in written production paradigms (Morgan & Wagers, 2018; Fadlon et al., 2019), resumption is thought to be a primarily spoken phenomenon in English (Polinsky et al., 2013). Furthermore, the fact that the base sentence remained on the screen for the duration of the trial in Exp. 2 might have encouraged participants to use unnatural strategies for completing the relative clause. For instance, the preponderance of *noun repetition* responses might reflect participants copying and pasting the final clause from the base sentence into the text box. We made several changes in Exp. 3 to increase its ecological validity and decrease the likelihood that participants used non-linguistic strategies like copy/pasting to perform the task.

Experiment 3: Spoken production

Experiment 3 was a production experiment which aimed to replicate the findings of Exp. 2 in a more ecologically valid paradigm. To boost ecological validity, a number of changes were made. First, responses were spoken rather than written. Second, once the participant read the base sentence, it was removed from the screen. The participant then performed a short arithmetic problem out loud to interrupt verbatim short-term memory and encourage more naturalistic, meaning-driven production rather than repetition of a surface-level representation (see Potter & Lombardi, 1990).

Method

Participants

48 undergraduates were recruited from the UC San Diego subject pool and received course credit for participation. This is the standard *N* for in-person studies in our lab, and was chosen prior to the beginning of subject running. Mean age was 20 (s.d. = 2); all were native

Table 7
Experiment 3 results. Model 1: The dependent variable was resumptive pronoun responses (coded as 1) vs. all other responses (coded as 0). Model 2: Excluding resumptive pronoun responses, 'other' responses (coded as 1) vs. gaps (coded as 0).

Model 1: RPs vs. gaps & 'other'	β	z	p	
Intercept (*weak island, distinct*)	−1.054	−5.724	< .001	***
SYNTAX (*non-island, distinct*)	−1.645	−6.032	< .001	***
SYNTAX (*strong island, distinct*)	1.215	6.264	< .001	***
LEXICALIZATION (*weak island, similar*)	−0.097	−0.481	.630	
Interaction (*non-island, similar*)	−0.337	−0.812	.417	
Interaction (*strong island, similar*)	−0.712	−2.580	.010	**
Model 2: 'Other' vs. gaps	β	z	p	
Intercept (*weak island, distinct*)	−0.434	−1.824	.068	.
SYNTAX (*non-island, distinct*)	0.286	1.304	.192	
SYNTAX (*strong island, distinct*)	1.041	3.824	< .001	***
LEXICALIZATION (*weak island, similar*)	1.252	5.300	< .001	***
Interaction (*non-island, similar*)	−0.185	−0.586	.558	
Interaction (*strong island, similar*)	0.234	0.589	.556	

monolingual speakers of American English.

Design

Stimuli were the same as in Exp. 2 (see Table 4).

Procedure

The experiment was programmed and run in PsychoPy (Peirce, 2007) and took roughly one hour to complete. Participants were told they would be performing a "paraphrase" task, completing new sentences (the preamble) on the basis of old ones (the base sentence). On each trial, participants first saw only the base sentence in red font on the screen, which they read out loud. At their own pace they then went on to the next screen, which displayed a simple addition problem. After completing this out loud, they saw the preamble, which they read aloud and completed based on the content of the base sentence.

They were instructed: "Read the red sentence out loud. When you think you can remember its meaning, push any button to go on to the next screen. Next, you will see a simple addition problem (e.g., 35 + 14). Say the answer ("49") out loud, and then push any key to go on. Finally, you will see the beginning of the new sentence. Begin reading it out loud. When you arrive at the ellipsis ('…'), continue speaking, using the information from the red sentence to finish this new one.".

Data coding and analysis

Data from a total of 1,728 critical trials were collected, of which 128 were excluded because participants skipped or otherwise made no effort to complete the sentence (e.g., "I forget"). Undergraduate assistants transcribed the first complete response, and ignored any subsequent attempts. The remaining 1,600 responses were coded according to the same rubric used in Exp. 2. A total of 468 were coded as *gap*; 426 as *resumptive pronoun*; 170 as *passives*; 58 as *noun repetition*; and 478 as *anomalous*. The analysis approach was identical to that in Exp. 2.

Results

The data are shown in Fig. 4 and model results in Table 7. Model 1 compared the rate of resumptive pronoun responses, coded as 1, to all other responses (including gap, passive, noun repetition, and anomalous responses), coded as 0. It converged with random intercepts for items and participants but no random slopes. Participants produced significantly more resumptive pronouns in weak islands than non-islands ($\beta = -1.645, z = -6.032, p < .001$), and even more in strong islands ($\beta = 1.215, z = 6.264, p < .001$), again replicating the findings of Morgan and Wagers (2018). There was no significant effect of LEXICALIZATION ($\beta = -0.097, z = -0.481, p = .630$), although a significant interaction term reflected the fact that, in strong islands, there were more resumptive pronouns in the *distinct* condition than *similar* ($\beta = -0.712, z = 2.580, p = .010$). Note that this effect is in the opposite direction of that predicted by the Acceptability Hypothesis.

The nested model, Model 2, excluded resumptive pronoun responses and compared gap responses, coded as 0, to all other remaining responses, coded as 1. It converged with random intercepts for items and participants but no random slopes. Results showed that there were fewer gap responses in strong islands than in weak islands ($\beta = 1.041, z = 3.824, p < .001$). Interestingly, despite the fact that LEXICALIZATION did not impact rates of resumption, it did affect gap production. Specifically, increasing lexical similarity decreased gap responses and increased 'other' responses ($\beta = 1.252, z = 5.300, p < .001$). (This pattern was also observed in Exp. 2, but it was only marginally significant.)

Discussion

Experiment 3 successfully replicated the important findings of Exp. 2 in a higher-powered, more ecologically valid study. Specifically, we again observed that non-islands received the fewest resumptive pronouns, then weak islands, and strong islands received the most. As for our critical manipulation, we again observed no main effect of LEXICALIZATION on resumptive pronoun production, consistent with the idea that acceptability does not drive resumptive pronoun production.

One interesting feature of the Exp. 3 data is the high number of anomalous responses – particularly in *similar* conditions. These responses consisted of ungrammatical or semantically incorrect responses. The high number of such responses relative to Exp. 2 likely reflects unique challenges of the spoken task. For instance, in the spoken task, only the first complete response was accepted, whereas in the written task participants could edit their responses before going onto the next trial. We suspect that the higher number of grammatical errors in Exp. 3 reflects the fact that errors are more likely to be corrected in writing. Similarly, in the written task participants saw the base sentence on the screen for the duration of the trial and could therefore refer back to it. However, in the spoken task the base sentence disappeared before the sentence completion portion of the trial. We suspect that many of the semantically incorrect responses (e.g., base: "The president wonders if the Saudi prince betrayed the contractor"; prompt: "That's the Saudi prince that the president wonders if"; response: "the Saudi prince had contracted") reflect incorrect memory of the base sentence.

Both Exps. 2 and 3 produced some other unexpected results. Some of these, such as the significant interaction term in Exp. 3, Model 1, were inconsistent across experiments and seemed likely to be spurious. Additionally, the marginal effect of LEXICALIZATION on non-island structures observed in Exp. 2 was not seen in Exp. 3 (where means patterned in the opposite direction), supporting our claim that the Exp. 2 difference was due to noise. However, the Model 2 effect of LEXICALIZATION appeared in both Exps. 2 (where it was marginal) and 3 (significant). To assess the reliability of these unexpected effects, we performed a final analysis.

Pooled Analysis of Experiments 2 & 3

We performed a pooled analysis, combining the data from Exps. 2 and 3 to ascertain whether certain effects that were either unexpected or inconsistent across the experiments would survive this higher-powered analysis. Specifically, we aimed to assess three results: (1) Exp. 2, Model 1 found a marginal interaction term reflecting the fact that in non-islands in that experiment were produced with more resumptive pronouns in the *similar* condition than in the *distinct* condition. If this effect is real, then acceptability may in fact drive resumptive pronoun production, but only in non-island structures. Given the peculiarity of this scenario and the marginal status of the effect, we suspect it to be spurious. (2) Experiment 3, Model 1 found a significant interaction reflecting the fact that in strong islands there were fewer resumptive pronouns in the *similar* condition than in the *distinct* condition. This

Table 8
Results of pooling data from Exps. 2 & 3. Model 1: The dependent variable was resumptive pronoun responses (coded as 1) vs. all other responses (coded as 0). Model 2: Excluding resumptive pronoun responses, 'other' responses (coded as 1) vs. gaps (coded as 0).

Model 1: RPs vs. gaps & 'other'	β	z	p	
Intercept (*weak island, distinct*)	−1.151	−7.219	< .001	***
SYNTAX (*non-island, distinct*)	−2.309	−6.994	< .001	***
SYNTAX (*strong island, distinct*)	1.368	7.422	< .001	***
LEXICALIZATION (*weak island, similar*)	−0.092	−0.556	.579	
Interaction (*non-island, similar*)	0.063	0.175	.861	
Interaction (*strong island, similar*)	−0.538	−2.374	.018	*
Model 2: 'Other' vs. gaps	β	z	p	
Intercept (*weak island, distinct*)	−1.130	−4.537	< .001	***
SYNTAX (*non-island, distinct*)	0.104	0.540	.590	
SYNTAX (*strong island, distinct*)	1.168	4.971	< .001	***
LEXICALIZATION (*weak island, similar*)	1.099	5.505	< .001	***
Interaction (*non-island, similar*)	0.023	0.087	.931	
Interaction (*strong island, similar*)	0.276	0.856	.392	

effect is in the opposite direction of that predicted by the Acceptability Hypothesis, and we think it is probably also spurious. (3) Finally, the nested models (Model 2) found an intriguing effect of LEXICALIZATION, whereby gaps were seen less when nouns were more *similar*. This effect was marginal in Exp. 2 but significant in Exp. 3. If real, then gaps trade off with other structures in response to increased lexical similarity.

Method

The analysis approach was the same as in Exps. 2 and 3. The models were reduced to facilitate convergence in the same way as previous models (see Exp. 1 *Method*).[6]

Results

Results appear in Table 8. Model 1 modeled resumption (1) vs. everything else (0) and converged with a random slope for SYNTAX that varied within participants and a random intercept for items. Results again showed a main effect of SYNTAX, whereby more resumptive pronouns are produced in weak islands than non-islands ($\beta = -2.309, z = 6.994, p < .001$), and more still in strong islands ($\beta = 1.368, z = 7.422, p < .001$), consistent with the main analyses in Exps. 2 and 3 (and previous literature).

Critically, Model 1 still finds no effect of LEXICALIZATION ($\beta = -0.092, z = -0.556, p = .579$), consistent with the prediction of the syntax model but not the acceptability model of resumptive pronoun production. The interaction term reflecting the effect of LEXICALIZATION in non-islands was not significant ($\beta = 0.063, z = 0.175, p = .861$), lending support to our suspicion that the marginal interaction term in Exp. 2 was likely spurious. However, the interaction term reflecting the effect of LEXICALIZATION in strong islands was significant ($\beta = -0.538, z = -2.374, p = .018$), although, again, this reflects an effect in the opposite direction of what would be predicted by an audience design account of resumption, suggesting that it too is likely spurious.

Model 2 converged with all three random intercepts, but no random slopes. Results again showed a strong effect of SYNTAX whereby strong islands elicited fewer gap responses than 'other' responses ($\beta = 1.168, z = 4.971, p < .001$). Interestingly, the effect of LEXICALIZATION was highly significant in the pooled model ($\beta = 1.099, z = 5.505, p < .001$), reflecting the fact that increased lexical similarity led to fewer gap production and more 'other' production.

Discussion

Several points can be made on the basis of the results of the pooled analysis. First, the unexpected interaction term that was marginally significant in Exp. 2 was not significant in the pooled analysis, suggesting that the marginal effect was likely spurious. Second, the unexpected significant interaction in Exp. 3 remained significant in the pooled analysis. However, even if this effect is real, it is in the opposite direction of what is predicted by the acceptability account of resumptive pronoun production, and thus could not serve as supporting evidence. We believe that this effect, too, is likely spurious, given that it only appeared in Exp. 3, is not predicted under any account, and is inconsistent with the results for the non-island and weak island structures (which we have no reason to expect to behave differently).

Third, the effect of SYNTAX on resumptive pronoun production is highly robust: non-islands receive fewer resumptive pronouns than weak islands, and weak islands receive fewer than strong islands.

Fourth, lexical similarity does not modulate rates of resumption. That is, in Exp. 2, Exp. 3, and the pooled analysis, the LEXICALIZATION term in Model 1 is not significant. This again comes with the standard caveat that null results cannot generally be interpreted as evidence for the lack of an effect. However, as argued above, it is reasonable to do so in this case.

Finally, LEXICALIZATION had an unexpected and intriguing effect on production. While it did not impact rates of resumption, it did lead a trade-off between gap responses and 'other' responses. We suspect this has to do with working memory demands, which are likely increased in *similar* conditions relative to *distinct* conditions, where nouns are less distinctive and therefore more difficult to retrieve from working memory (Lewis, 1996).

General Discussion

We began by pointing out a contradiction in two prominent psycholinguistic literatures. On the one hand, research on audience design has produced mounting evidence that producers do not choose particular syntactic structures to benefit their interlocutors (Arnold et al., 2004; Ferreira & Dell, 2000; Jaeger, 2010; Ferreira & Schotter, 2013). Indeed, evidence for syntactic audience design has been so elusive that its absence is reflected in the basic architecture of relevant production models (Levelt, 1993; Bock & Levelt, 1994; Ferreira et al., 2018; Garrett, 1975). If these models are correct, then not only do producers not choose syntactic structures on the basis of the case-by-case needs of their interlocutors, it is in fact impossible for them to do so directly.

On the other hand, in the literature on long-distance dependencies, audience design plays a central role in many theories of why English speakers produce resumptive pronouns. Many of these theories hold that resumptive pronouns serve to improve acceptability – what we have referred to as the Acceptability Hypothesis. Indeed, recent evidence appears to support this idea: resumptive pronoun production rates are strongly negatively correlated with gap acceptability (Morgan & Wagers, 2018). Resumption, then, may constitute a Hail Mary in the search for syntactic audience design.

However, the finding of a correlation between acceptability and resumptive pronoun production does not necessarily imply audience design. Indeed, this assumes an underlying causal relationship, and correlation of course does not imply causation. It is also possible that some latent variable (e.g., processing difficulty) causally impacts both acceptability and production, leading to the observed correlation.

[6] We initially attempted to model the pooled data with an additional fixed effect, EXPERIMENT (i.e., *written* or *spoken*), as well as with its interactions with SYNTAX and LEXICALIZATION and a random slope for EXPERIMENT that varied within items. However, this model would not converge with any random effects structure, nor would subsequent models with fewer interaction terms (e.g., without EXPERIMENT × SYNTAX). As the goal of this analysis is to determine whether differences thought to be spurious in Exps. 2 and 3 persisted in a higher-powered analysis, not to assess differences between the experiments, we opted to remove the EXPERIMENT term altogether to report the results of models with fuller random effects structures.

Our approach to resolving the discrepancy between the audience design literature and the long-distance dependency literature was to pit the Acceptability Hypothesis against this latent variable hypothesis. Specifically, if some construct other than acceptability is the true underlying cause of resumption, insofar as that construct is separable from acceptability it should be possible to manipulate acceptability without changing resumptive pronoun production rates. But if the Acceptability Hypothesis is correct, then any change to acceptability should lead to corresponding changes in production.

Morgan and Wagers (2018) observed the correlation between acceptability and production across syntactic structures, indicating that whatever the true cause of resumption is, it varies with syntax. We therefore disentangled syntax and acceptability by manipulating lexical similarity. If unacceptability causes resumptive pronoun production, speakers should produce more resumptive pronouns in sentences with more similar nouns (which are less acceptable) than more distinct nouns (which are more acceptable), even when structure is held constant. But if some other property of syntactic structure is to blame, then it might be possible to manipulate acceptability without causing a change in production.

Experiment 1 validated lexical similarity as a tool for manipulating sentence acceptability. Specifically, we manipulated sentences' lexicalizations and their syntactic structures, and found that both had the expected effects on acceptability. Experiment 2 was a written production task which elicited sentences with gaps and resumptive pronouns, with the same manipulations as Exp. 1. Results showed a clear effect of syntax, but no effect of lexicalization, which was inconsistent with the Acceptability Hypothesis. Experiment 3 successfully replicated the results of Exp. 2, but with higher power and better ecological validity. A final, pooled analysis of the Exp. 2 and Exp. 3 data produced the same result. The evidence, then, is against the Acceptability Hypothesis.

Interpreting null results

One potential weakness of this conclusion is that it relies heavily on a null result. In general, null results cannot be taken as evidence for the absence of an effect. The primary risk is that there is in fact an underlying effect, but it is too small to be detected given the statistical power. However, as mentioned above, this is unlikely to be a risk in the present case.

To understand why, consider the results of Exp. 1 (Table 2). The LEXICALIZATION and SYNTAX manipulations impacted acceptability to comparable degrees (compare the Cohen's *d* values for LEXICALIZATION, 0.305 and 0.581, to those for SYNTAX, 0.327, 0.198, and 0.508). If acceptability is the cause of resumptive pronoun production, then we should expect to see both manipulations result in production differences of comparable size.[7] Instead, while the effect of SYNTAX on resumptive pronoun production was extremely robust ($p < .001$ in the pooled analysis), neither Exp. 2 nor 3 found any effect of lexicalization. Given the robustness of the SYNTAX effect and the successful replication of this pattern in Exp. 3, the null result is not likely to reflect a lack of power. In this case, we believe it is reasonable to directly interpret the null result, and conclude that unacceptability is not the cause of resumptive pronoun production.

Does audience design require an audience?

A potential shortcoming of our study has to do with the nature of audience design effects. On the basis of the absence of a lexical similarity effect in Exps. 2 & 3, we have argued resumptive pronoun production is not the result of audience design. However, participants in our production studies had no clear "audience." Experiments 2 & 3 were described to participants as "paraphrase" tasks; participants were not speaking *to* anyone in particular, nor were they instructed that their utterances should have any communicative goal. The argument against audience design, then, relies on the assumption that audience design should occur even in the absence of an audience.

However, there is some experimental evidence that for audience design, the presence of an addressee is unlikely to be relevant. In particular, Ferreira and Dell (2000) present six experiments testing whether the use of the optional complementizer *that* (as in *The coach knew (that) you missed practice*) was better explained by production pressures or an audience design account. Whereas the first five experiments used memory-recall paradigms, which are no more audience-motivated than the present experiments, in the final experiment participants spoke directly to a live, experimentally-naive interlocutor, and were instructed to make their utterances as easy to understand as possible. The results patterned with those of the previous five experiments, favoring the production-based account over an audience design account.

Furthermore, it is not clear that a communicative goal is necessary for audience design to occur – or at least for a particular kind of audience design. Dell and Brown (1991) argue for a distinction between two types of audience design: *particular-listener adaptation*, where producers formulate their utterances so as to meet the needs of a particular comprehender, and *generic-listener adaptation*, where producers formulate utterances so as to be easy to comprehend in general.

While particular-listener adaptation requires that there is a specific interlocutor who will need to interpret the producer's utterances, which was not the case in our experiments, generic-listener adaptation is taken to be a property of language production in general, either because it is an inherent part of the language system as it evolved, or because the system learns over time which kinds of utterances are the most effective for successfully conveying a particular message (see, e.g., Ferreira, 2019).

In our production experiments, any audience design effects would thus have fallen under the umbrella of generic-listener adaptation. A conservative statement of our findings, then, is that English speakers do not produce resumptive pronouns to make their utterances more generally comprehensible. To make the broader claim that resumption is not the result of audience design at all will require further research to determine whether production patterns change when producers have a clear communicative goal.

Acceptability and acceptability judgments

We have so far taken for granted the existence of *acceptability* as a psychological construct, and of *acceptability ratings* as a reasonable metric of this construct. However, acceptability is in fact notoriously difficult to define, and how speakers map from acceptability to acceptability ratings remains an open question.

The notion of acceptability became especially prominent in the 21st Century with renewed interest in formal syntactic modeling, such as that pioneered by Noam Chomsky. Acceptability was taken to be a reflection of whether a sentence was underlyingly grammatical or ungrammatical, and is therefore the primary variable of interest in research that aims to characterize or model the syntax of a language.

As the use of acceptability ratings became widespread in experimental syntax research, so did research into how exactly to interpret this measure (Sprouse, 2007; Schütze & Sprouse, 2013). It became clear early on that acceptability ratings do not directly reflect grammaticality, or at least not exclusively. While grammaticality is generally taken to be binary – a speaker either has a representation of a particular syntactic structure (i.e., it is grammatical) or they do not (it is ungrammatical) – acceptability ratings appear to be gradient, reflecting the contributions

[7] Note that grammaticality also seems to independently boost rates of resumptive pronouns in a structure, accounting for the fact that, more than can be accounted for by acceptability alone, resumptive pronouns are more common in islands than non-islands (Morgan & Wagers, 2018). For simplicity, we ignore this issue here. It is safe to do so given that our argument holds even if we limit the scope to just the two island conditions, which do not differ in grammaticality.

of a number of other underlying properties of sentences including structural frequency, processing difficulty, semantic coherence, etc. (Liu, Ryskin, Futrell, & Gibson, 2022; Schütze & Sprouse, 2013; Sprouse, 2007).

A reasonable concern could be that, acceptability being so far from clearly defined, there might be room for alternative interpretations of our results that do not bear on audience design. However we think this is unlikely. Specifically, for the logic of our conclusions to hold, two important conditions on acceptability judgments must be met. First, whatever acceptability ratings measure, they must do so *reliably*. That is, we should be reasonably confident that participants are all reporting a metric of the same construct. Evidence for this comes from high inter-study reliability (Sprouse & Almeida, 2017; Sprouse, Schütze, & Almeida, 2013), even when studies use different paradigms (e.g., Likert rating vs. magnitude estimation; although see Sprouse, 2011). (See also the above discussion in the Introduction on the relatively nuanced pattern in acceptability judgments for resumptive pronouns that has replicated across several studies.).

The second important condition has to do with the notion of *validity* – that acceptability ratings measure the construct we want it to. While acceptability is hard to define, what is critical for our purposes is that it is not a property of production, but of comprehension (or, more precisely in this context, of the speaker's assessment of how their utterance will be comprehended). If this is the case, then any production outcomes that depend on acceptability might be said to reflect audience design.

Indeed, in cases where comprehension and production patterns diverge, acceptability ratings seem to track with comprehension. Coincidentally, some of the best evidence for this comes from previous findings in the resumptive pronoun literature. For instance, in structures where resumptive pronouns are interpreted incorrectly yet commonly produced, acceptability ratings are low, patterning with the interpretation data and not the production data (Morgan et al., 2020; see also other dissociations discussed in Han et al., 2012; Keffala & Goodall, 2011; Polinsky et al., 2013; Morgan & Wagers, 2018).

Overall, then, the evidence suggests that acceptability judgments are a reasonable measure for addressing questions about audience design like the one we have asked in this study.

Other effects of lexical similarity on production

A point that we have glossed over is the finding that increased lexical similarity led to increased passives, noun repetitions, and anomalous responses. We have argued that such an effect for resumptive pronouns would have been evidence of syntactic audience design, but we have refrained from making the same argument for these other structures. This is because we believe that there are better explanations for this pattern of data.

Specifically, it has been suggested that the mechanism by which lexical similarity impacts acceptability is working memory (Bever, 1974; Lewis, 1996). This would be an example of *similarity-based interference*, an effect whereby the more similar the items held in working memory, the less accurately they are recalled (Shulman, 1971; Waugh & Norman, 1965). If this is correct, then our lexicalization manipulation was fundamentally a memory load manipulation, and we should expect to see signatures of higher working memory burden on production outcomes when lexical similarity was higher.

We believe this accounts for the increased rates of "other" response types. For instance, one consequence of increasing the working memory load during speech production is an increase in speech errors (Ivanova & Ferreira, 2019), including syntactically and/or semantically incorrect responses such as those we categorized as noun repetition and anomalous responses. Indeed, it would be hard to make an audience design case for either of these response types, as it would rely on arguing that speakers choose incorrect responses to help comprehenders. We think it is more likely that the higher rates of these responses in *similar* conditions reflect a breakdown in production due to increased difficulty.

As for passive responses, these have been argued to be a strategy that producers use to reduce difficulty during the production of particularly processing-intensive structures. Indeed, For instance, using stimuli similar to ours (where a gap appears in an embedded object position) Gennari, Mirković, and MacDonald (2012) showed that speakers tend to produce passives when doing so would decrease similarity-based interference among the various nouns in the sentence. (See also Scontras, Badecker, Shank, Lim, & Fedorenko, 2015 for another production-based account of this phenomenon. Notably, both of these explanations are production-based and do not involve a role for audience design.).

It is true, however, that what is easier for producers is often easier for comprehenders, making it hard to disentangle accounts based on audience design from those rooted in other production pressures. Indeed, when lexical similarity is high, passivization in these types of sentences also facilitates comprehension (Gordon, Hendrick, & Johnson, 2001; King & Just, 1991; MacWhinney & Pléh, 1988; Traxler, Morris, & Seely, 2002). It may therefore still be possible to make a case for passivization as a form of audience design, although we consider this unlikely in light of the existence of compelling difficulty-based accounts, as well as the broader lack of evidence for syntactic audience design. This is ultimately an empirical question, and a good goal for future research will be to dissociate whether speakers' behaviors are driven by what is difficult for producers versus for comprehenders.

Why do speakers produce resumptive pronouns?

If not acceptability, then what *does* lead speakers to produce resumptive pronouns? There are a number of other hypotheses in the literature. For instance, Beltrama and Xiang (2016) hypothesized that *comprehensibility* plays a role, and showed that people rate sentences with resumptive pronouns as more comprehensible than sentences with gaps. However, perhaps even more than with acceptability, a comprehensibility explanation implies a central role for syntactic audience design. As discussed above, existing evidence (or, more specifically, the lack thereof) indicates that speakers do not choose their syntax for the benefit of the comprehender. Furthermore, recent work indicates that even in sentences where speakers produce more resumptive pronouns than gaps, comprehenders are less correct in their interpretations of resumptive pronouns than gaps (Morgan et al., 2020). This of course does not preclude the possibility that speakers *attempt* to make their utterances more comprehensible with resumptive pronouns, even if such efforts are in vain. But taken together, these facts suggest the likelihood of a comprehensibility-based explanation is low.

Processing difficulty is another commonly-cited candidate (Prince, 1990; Alexopoulou & Keller, 2007; Polinsky et al., 2013; Goodall, 2015; Meltzer-Asscher, 2021). It fits nicely with the pattern of data observed in Morgan and Wagers (2018) in that it varies across structures (Staum Casasanto et al., 2010), correlates with acceptability (Hofmeister et al., 2013; Hofmeister et al., 2014), and has been shown to be a relevant factor in grammatical resumption in Hebrew (Fadlon et al., 2019; Keshev & Meltzer-Asscher, 2017). It also has the benefit of not necessarily implicating audience design. If the causal role can be assigned to a construct like processing difficulty rather than acceptability, it may be straightforward to resolve the contradiction between the audience design literature and the long-distance dependency literature.

However, as far as explanations go, a processing difficulty account is somewhat vague. As a construct, it probably reflects a number of underlying properties like complexity and working memory load. While we remain largely agnostic as to which (if any) of these is the underlying cause of resumptive pronoun production, some likely characteristics of the true cause can be deduced by considering the data. These may be helpful in narrowing down the possibility space.

First, because resumptive pronouns are produced at different rates in different structures, the relevant construct should be a property of particular syntactic structures. Second, because rates of resumptive pronouns correlate across structures with acceptability, this construct

should also correlate with structural acceptability.

Third, if our conclusion that lexical similarity does not impact the production of resumptive pronouns is correct, then lexical similarity does not affect this construct. This observation may be particularly important, as lexical similarity is linked to other constructs that researchers think may cause resumptive pronoun production.

Our findings might therefore be re-stated as evidence against not only an acceptability account of resumption, but also against processing difficulty accounts that rely on a working-memory mechanism. These accounts tend to be production-based (Fadlon et al., 2019; Ariel, 1999; Polinsky et al., 2013; but see Chacón, 2019 for a comprehension-based working memory account), arguing that resumptive pronouns are the result of overwhelming the producer's working memory system. If our lexical similarity manipulation is indeed a working memory manipulation, then our results cast doubt on the viability of working-memory based theories as well.

Another intriguing possibility is that our characterization of acceptability as a property of whole utterances is wrong. It may instead be the case that that different aspects of a sentence – its semantics, its lexicalization, its syntactic structure – each come with their own independent degree of acceptability, and these combine to form a global degree of acceptability. Such an account may provide a way to salvage the Acceptability Hypothesis, as it would remain possible that *structural acceptability* is what drives resumptive pronoun production, not lexical acceptability. This would have important implications for models of speech production. Specifically, it would implicate a stage in the production of long-distance dependencies during which the entire delexicalized structure is planned, assessed for acceptability, and a decision about whether to produce a gap or resumptive pronoun is made.

Finally, we wish to highlight the possibility of another brand of explanation that does not rely on a particular trigger like acceptability or working memory. Specifically, recent theoretical work suggests that structural priming effects may reflect a slow-acting learning mechanism (Chang, Dell, & Bock, 2006; Fine & Florian Jaeger, 2013; Ferreira, 2019). Over time, and with lots of exposure to, for instance, passives and actives, the accessibility of these structural alternates comes to mirror their relative frequencies in the environment. (Interestingly, there is a degree of circularity here in that, within a linguistic community, the structures a speaker produces are determined by those structures' environmental frequencies, and simultaneously impact their own environmental frequencies.) During speech production, particular structures are chosen stochastically, but with probabilities corresponding to their accessibility.

A similar account could hold for long-distance dependencies: Over time, and with lots of exposure to long-distance dependencies in various structures (including islands and non-islands), the system eventually learns the structure-specific frequency of gaps and/or resumptive pronouns in the environment. This translates into differing degrees of accessibility for resumptive pronouns across structures. During production, then, speakers choose resumptive pronouns stochastically, but with probability weighted according to their likelihood of appearing in that syntactic structure in the environment.

A potential shortcoming of this account is that it treats gaps and resumptive pronouns as if they were simply structural alternates – that is, both grammatical ways of encoding the same message, similar to the active/passive alternation. However, researchers tend to agree that resumptive pronouns are not in fact grammatical in English (although this is not quite consensus; see, e.g., Creswell, 2002; Goodall, 2017; and Asudeh, 2012, who argues that even though resumptive pronouns are ungrammatical at the level of the whole sentence, the system may in fact consider them grammatical on a more local scale). We suspect that a more nuanced look at the relationship between speakers' history of exposure to resumptive pronouns and their propensity to produce them (as in a standard priming experiment) stands to shed significant light on the question of how resumptive pronouns are represented and why they appear in speech.

Conclusion

In the literature on audience design, there is general agreement that producers do not choose particular syntactic structures for the benefit of their interlocutors. However, exactly this type of mechanism is at the heart of many models of English resumption. For instance, some models propose that speakers produce resumptive pronouns to improve the acceptability of their utterances. Indeed, a recent finding that resumptive pronoun production correlates with acceptability seems to support this idea.

In three experiments, this paper investigated this correlation. We manipulated sentences' acceptability independent of their syntactic structure, and found no effect of this manipulation on resumptive pronoun production, indicating that low acceptability does not cause speakers to produce resumptive pronouns. Resumption, then, is another example of a case where syntactic audience design seems likely but does not in fact manifest. We take this as further evidence in support of the idea that syntactic audience design is in fact not possible.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Supplementary Analyses

The use of local person pronouns to create mutual distinctiveness among NPs in our stimuli may have introduced unintended distinctions in information structure between the *similar* and *distinct* conditions (see Warren & Gibson, 2005 for evidence that indexical pronouns are processed differently than other types of nominals). To determine whether our findings might have been impacted by these unintended differences, we re-ran the pooled analysis of Experiments 2 and 3, but this time using only those items that did not use local person pronouns in the *distinct* condition. The *non-island, distinct* stimuli all included one local person pronoun (these were formed by using one of each of the four types of nominals), so these were altogether removed the data. The resulting dataset, shown in Fig. A.1, came from 14 item sets in a 2 × 2 design: SYNTAX (*weak island* or *strong island*) and LEXICALIZATION (*similar* or *distinct*).

The results of the analysis, reported in Table A.1, are slightly different from those reported in the pooled analysis (Table 8), though not in a way that impacts our overall conclusion. There are two minor differences: First, the interaction in Model 1, which was significant in the pooled analysis, is not significant in the supplemental analysis (lending further support to our argument that this effect was spurious; $\beta = 0.202, z = 0.490, p = .624$). Second, the interaction term in Model 2, which was not significant in the pooled analysis, was marginally significant in the supplemental analysis ($\beta = 1.644, z = 1.943, p = .052$), reflecting the fact that the increase in 'other' structures relative to gaps in *similar* conditions was larger in strong islands than in weak islands. (Note that this difference-of-differences is clearer in logits than in the proportions shown in Fig. A.1).

The most important difference is in the Model 1 effect of LEXICALIZATION, which was not significant in the pooled analysis, but marginally significant in the supplemental analysis ($\beta = -.613, z = -1.733, p = .083$). Critically, however, the effect is *negative* – the opposite direction predicted by the

Acceptability Hypothesis – so even if it is real, it does not constitute support an audience design account of resumption.

Thus, even when deconfounding the acceptability manipulation from whatever differences come with the use of local person pronouns, the data still support the conclusion that resumptive pronoun production in English is not the result of audience design.

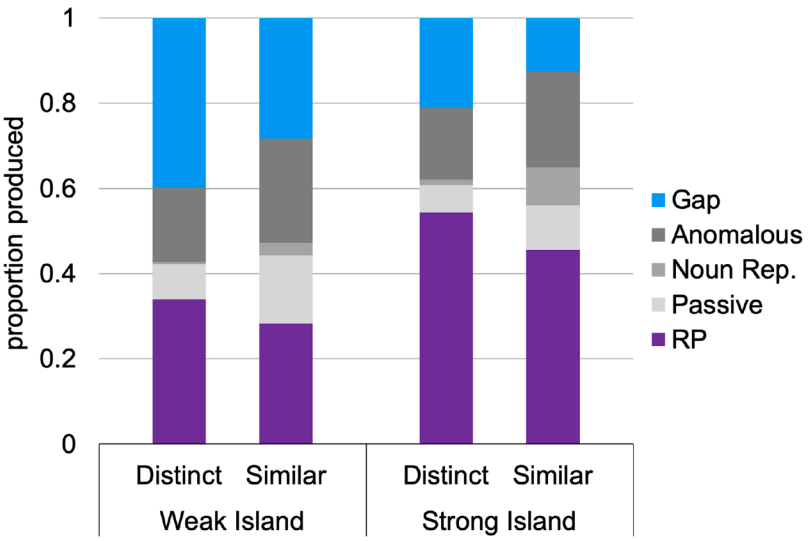

Fig. A.1. Data for supplemental analysis: Proportions of each coded response type per experimental condition in pooled data from Exps. 2 & 3, excluding items sets that used local person pronouns in the *distinct* condition.

Table A.1
Results of pooling data from Exps. 2 & 3 and excluding item sets with local person pronouns. Model 1: The dependent variable was resumptive pronoun responses (coded as 1) vs. all other responses (coded as 0). Model 2: Excluding resumptive pronoun responses, 'other' responses (coded as 1) vs. gaps (coded as 0).

Model 1: RPs vs. gaps & 'other'	*β*	*z*	*p*	
Intercept (*weak island, distinct*)	−0.801	−3.243	.001	**
SYNTAX (*strong island, distinct*)	1.011	3.440	< .001	***
LEXICALIZATION (*weak island, similar*)	−0.613	−1.733	.083	.
Interaction (*strong island, similar*)	0.202	0.490	.624	
Model 2: 'Other' vs. gaps	*β*	*z*	*p*	
Intercept (*weak island, distinct*)	−0.532	−1.058	.290	
SYNTAX (*strong island, distinct*)	1.138	1.896	.058	.
LEXICALIZATION (*weak island, similar*)	0.099	0.156	.876	
Interaction (*strong island, similar*)	1.644	1.943	.052	.

References

Ackerman, L., Frazier, M., & Yoshida, M. (2018). Resumptive pronouns can ameliorate illicit island extractions. *Linguistic Inquiry, 49*(4), 847–859. https://doi.org/10.1162/ling_a_00291

Alexopoulou, T., & Keller, F. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, 110–160.

Aoun, J., Choueiri, L., & Hornstein, N. (2001). Resumption, movement, and derivational economy. *Linguistic Inquiry, 32*(3), 371–403.

Ariel, M. (1999). Cognitive universals and linguistic conventions: The case of resumptive pronouns. Studies in Language. International Journal sponsored by the Foundation. *Foundations of Language, 23*(2), 217–269.

Arnold, J. E., Wasow, T., Asudeh, A., & Alrenga, P. (2004). Avoiding attachment ambiguities: The role of constituent ordering. *Journal of Memory and Language, 51*(1), 55–70.

Asudeh, A. (2011). Local grammaticality in syntactic production. *Language from a Cognitive Perspective*, 51–79.

Asudeh, A. (2012). *The logic of pronominal resumption*. Oxford University Press.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. arXiv preprint arXiv:1406.5823.

Beltrama, A., & Xiang, M. (2016). Unacceptable but comprehensible: The facilitation effect of resumptive pronouns. *Glossa, 1*(1), 1.

Bennett, R. (2009). English resumptive pronouns and the highest-subject restriction: A corpus study. Trilateral (TREND) Linguistics Weekend, UC Santa Cruz.

Bever, T.G. (1970). The cognitive basis for linguistic structures. Cognition and the development of language.

Bever, T. G. (1974). The ascent of the specious, or there's a lot we don't know about mirrors. *Explaining linguistic phenomena*, 173–200.

Bock, J. K. (1982). Toward a cognitive psychology of syntax: Information processing contributions to sentence formulation. *Psychological Review, 89*(1), 1.

Bock, J. K., & Levelt, W. J. (1994). *Language production: Grammatical encoding*. Academic Press.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(6), 1482.

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language, 89*, 68–86.

Chacón, D.A.(2019). Minding the gap?: Mechanisms underlying resumption in English. Glossa: a journal of general linguistics 4(1).

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological review, 113*(2), 234.

Clemens, L.E., Morgan, A., Polinsky, M., & Xiang, M.(2012). Listening to resumptives: An auditory experiment. In poster presented at the 25th annual CUNY Conference on Human Sentence Processing, New York.

Creswell, C. (2002). Resumptive pronouns, wh-island violations, and sentence production. In Proceedings of the sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (tag+ 6) (pp. 40–47).

Dell, G. S., & Brown, P. M. (1991). Mechanisms for listener-adaptation in language production: Limiting the role of the model of the listener. In *Bridges between psychology and linguistics* (pp. 117–142). Psychology Press.

Dickey, M. W. (1996). Constraints on the sentence processor and the distribution of resumptive pronouns. *Linguistics in the Laboratory, 19*, 157–192.

Drummond, A.(2013). Ibex farm. Online server: http://spellout.net/ibexfarm.

Elsness, J. (1984). *That* or zero? A look at the choice of object clause connective in a corpus of American English. *Engl. Stud., 65*, 519–533.

Engdahl, E. (1985). Parasitic gaps, resumptive pronouns, and subject extractions. *Language, 23*(1). https://doi.org/10.1515/ling.1985.23.1.3
Fadlon, J., Morgan, A. M., Meltzer-Asscher, A., & Ferreira, V. S. (2019). It depends: Optionality in the production of filler-gap dependencies. *Journal of Memory and Language, 106*, 40–76.
Ferreira, F., & Clifton, C., Jr (1986). The independence of syntactic processing. *Journal of Memory and Language, 25*(3), 348–368.
Ferreira, F., & Swets, B.(2005). The production and comprehension of resumptive pronouns in relative clause island contexts. Twenty-first Century Psycholinguistics: Four Cornerstones, 263–278.
Ferreira, V. S. (2019). A mechanistic framework for explaining audience design in language production. *Annual Review of Psychology, 70*, 29–51.
Ferreira, V. S., & Dell, G. S. (2000). Effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology, 40*(4), 296–340.
Ferreira, V. S., Morgan, A. M., & Slevc, L. R. (2018). Grammatical encoding. *The Oxford Handbook of Psycholinguistics*. https://doi.org/10.1093/oxfordhb/9780198786825.013.18
Ferreira, V. S., & Schotter, E. R. (2013). Do verb bias effects on sentence production reflect sensitivity to comprehension or production factors? *Quarterly Journal of Experimental Psychology, 66*(8), 1548–1571.
Fine, A. B., & Florian Jaeger, T. (2013). Evidence for implicit learning in syntactic comprehension. *Cognitive Science, 37*(3), 578–591.
Garrett, M. F. (1975). *The analysis of sentence production. In Psychology of learning and motivation* (Vol. 9, pp. 133–177). Elsevier.
Gennari, S. P., Mirković, J., & MacDonald, M. C. (2012). Animacy and competition in relative clause production: A cross-linguistic investigation. *Cognitive Psychology, 65*(2), 141–176.
Goodall, G. (2015). The D-linking effect on extraction from islands and non-islands. *Frontiers in Psychology, 5*, 1493.
Goodall, G. (2017). Referentiality and resumption in wh-dependencies. Asking the right questions: Essays in honor of Sandra Chung, 65–80.
Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(6), 1411.
Hammerly, C. (2019). The pronoun which comprehenders who process it in islands derive a benefit. *Linguistic Inquiry*, 1–21. https://doi.org/10.1162/ling_a_00422
Han, C-h., Elouazizi, N., Galeano, C., Görgülü, E., Hedberg, N., Hinnell, J.... Kirby, S. (2012). Processing strategies and resumptive pronouns in english. In Proceedings of the 30th West Coast Conference on Formal Linguistics (pp. 153–161).
Heestand, D., Xiang, M., & Polinsky, M. (2011). Resumption still does not rescue islands. *Linguistic Inquiry, 42*(1), 138–152.
Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19–24.
Hofmeister, P., Casasanto, L. S., & Sag, I. A. (2014). Processing effects in linguistic judgment data: (super-)additivity and reading span scores. *Language and Cognition, 6*(1), 111–145.
Hofmeister, P., Jaeger, T. F., Arnon, I., Sag, I. A., & Snider, N. (2013). The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes, 28*(1–2), 48–87.
Hofmeister, P., & Norcliffe, E. (2013). *Does resumption facilitate sentence comprehension? In The core and the periphery: Data-driven perspectives on syntax inspired by Ivan A. Sag* (pp. 225–246). CSLI Publications.
Ivanova, I., & Ferreira, V. S. (2019). The role of working memory for syntactic formulation in language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(10), 1791.
Jaeger, T. F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive psychology, 61*(1), 23–62.
Jaeger, T. F., & Ferreira, V. (2013). Seeking predictions from a predictive framework. *The Behavioral and Brain Sciences, 36*(4), 359.
Keffala, B., & Goodall, G. (2011). Do resumptive pronouns ever rescue illicit gaps in English. In poster presented at CUNY 2011 Conference on Human Sentence processing.
Keshev, M., & Meltzer-Asscher, A. (2017). Active dependency formation in islands: How grammatical resumption affects sentence processing. *Language, 93*(3), 549–568.
King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of memory and language, 30*(5), 580–602.
Koopman, H. (1983). Control from comp and comparative syntax. *The linguistic review, 2*(4), 365–391.
Kroch, A.S. (1981). On the role of resumptive pronouns in amnestying island constraint violations. In Papers from the Regional Meeting of the Chicago Linguistics Society. Chicago, Ill. (pp. 125–135).
Kurumada, C., & Jaeger, T. F. (2015). Communicative efficiency in language production: Optional case-marking in Japanese. *Journal of Memory and Language, 83*, 152–178.
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13
Levelt, W. J. (1993). *Speaking: From intention to articulation* (Vol. 1). MIT Press.
Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25*(1), 93–115.
Liu, Y., Ryskin, R., Futrell, R., & Gibson, E. (2022). A verb-frame frequency account of constraints on long-distance dependencies in English. *Cognition, 222*, 104902.
Lombard, E. (1911). Le signe de l'elevation de la voix. Ann. Mal. de L'Oreille et du Larynx, 101–119.
MacWhinney, B., & Pléh, C. (1988). The processing of restrictive relative clauses in hungarian. *Cognition, 29*(2), 95–141.
McCloskey, J. (1990). Resumptive pronouns, Ā-binding and levels of representation in Irish. In R. Hendrick (Ed.), Syntax of the modern Celtic languages (Vol. 23, pp. 199–248). New York and San Diego: Academic Press. (Republished in Rouveret (2011), pp 65–119.).
McDaniel, D., & Cowart, W. (1999). Experimental evidence for a minimalist account of English resumptive pronouns. *Cognition, 70*(2), B15–B24.
Meltzer-Asscher, A. (2021). Resumptive pronouns in language comprehension and production. *Annual Review of Linguistics, 7*, 177–194.
Morgan, A. M. (2022). The that-trace effect and island boundary-gap effect are the same: Demonstrating equivalence with null hypothesis significance testing and psychometrics. *Glossa Psycholinguistics, 1*, 1.
Morgan, A. M., von der Malsburg, T., Ferreira, V. S., & Wittenberg, E. (2020). Shared syntax between comprehension and production: Multi-paradigm evidence that resumptive pronouns hinder comprehension. *Cognition, 205*, 104417.
Morgan, A. M., & Wagers, M. W. (2018). English resumptive pronouns are more common where gaps are less acceptable. *Linguistic Inquiry, 49*(4), 861–876.
Pan, V. J. (2016). *Resumptivity in Mandarin Chinese*. De Gruyter Mouton.
Peirce, J. W. (2007). Psychopy-psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1–2), 8–13.
Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences, 27*(2), 169–190.
Polinsky, M., Clemens, L.E., Morgan, A.M., Xiang, M., & Heestand, D.(2013). Resumption in English. Experimental syntax and island effects 341.
Potter, M. C., & Lombardi, L. (1990). Regeneration in the short-term recall of sentences. *Journal of Memory and Language, 29*(6), 633–654.
Prince, E.F.(1990). Syntax and discourse: A look at resumptive pronouns. In Annual Meeting of the Berkeley Linguistics Society (Vol. 16, pp. 482–497).
R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/.
Schütze, C.T., & Sprouse, J.(2013). Judgment data. Research methods in linguistics 27–50.
Scontras, G., Badecker, W., Shank, L., Lim, E., & Fedorenko, E. (2015). Syntactic complexity effects in sentence production. *Cognitive science, 39*(3), 559–583.
Shlonsky, U. (1992). Resumptive pronouns as a last resort. *Linguistic inquiry, 23*(3), 443–468.
Shulman, H. G. (1971). Similarity effects in short-term memory. *Psychological Bulletin, 75*(6), 399.
Snow, C. E. (1972). Mothers' speech to children learning language. *Child development*, 549–565.
Sprouse, J. (2007). Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics, 1*, 123–134.
Sprouse, J. (2011). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, 274–288.
Sprouse, J., & Almeida, D. (2017). Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences, 40*, e311.
Sprouse, J., Schütze, C. T., & Almeida, D. (2013). A comparison of informal and formal acceptability judgments using a random sample from linguistic inquiry 2001–2010. *Lingua, 134*, 219–248.
Staum Casasanto, L., Hofmeister, P., & Sag, I.A. (2010). Understanding acceptability judgments: Additivity and working memory effects. In Proceedings of the 2010 Annual Meeting of the Cognitive Science Society, Portland Ore. USA.
Temperley, D. (2003). Ambiguity avoidance in English relative clauses. *Language, 79*, 464–484.
Traxler, M. J., Morris, R. K., & Seely, R. E. (2002). Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language, 47*(1), 69–90.
Trueswell, J. C., Tanenhaus, M. K., & Kello, C. (1993). Verb-specific constraints in sentence processing: separating effects of lexical preference from garden-paths. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(3), 528.
Warren, T., & Gibson, E. (2005). Effects of np type in reading cleft sentences in english. *Language and Cognitive Processes, 20*(6), 751–767.
Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological review, 72*(2), 89.
Zenker, F., & Schwartz, B.D. (2021). Resumptive pronouns facilitate processing of long-distance relative clause dependencies in second language English. In Proceedings of the Linguistic Society of America (Vol. 6, pp. 325–339). doi:10.3765/plsa.v6i1.4972.
Zollinger, S. A., & Brumm, H. (2011). The Lombard Effect. *Current Biology, 21*(16), R614–R615.