The role of referent predictability in pronoun production: Insights from a Bayesian meta-analysis

Xixian Liao[1,2], Thomas Brochhagen[2], Gemma Boleda[2,3], and Laia Mayol[2]

[1] *Language Technologies Unit, Barcelona Supercomputing Center, Barcelona, Spain*
[2] *Department of Translation and Language Sciences, Universitat Pompeu Fabra, Barcelona, Spain*
[3] *Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain*

The role of referent predictability in pronoun production: Insights from a Bayesian meta-analysis

Abstract

While it is known that speakers tend to use more reduced forms for more predictable words or phrases, it is unclear whether the same happens at the referential level: The influence of referent predictability on pronoun production remains a contentious issue, with divergent findings reported in the literature. To address this inconsistency, we carried out a Bayesian meta-analysis of the current literature on the relationship between referent predictability and pronoun production. Our meta-analysis covers 20 primary peer-reviewed studies, encompassing 26 experiments across 8 languages. We find stronger evidence for a small positive effect of referent predictability on pronoun usage, as opposed to the alternative hypothesis of no effect or a negative effect. As the first comprehensive synthesis of available evidence on this topic, our study offers insights into pronoun production and identifies promising avenues for future research: focus on typologically diverse languages, on conditions where a variety of referring expressions are expected or where the effect of predictability is more likely to appear, among others. Finally, we also advocate for the use of meta-analysis as a tool for theoretical linguistics.*

Keywords: meta-analysis, pronoun production, predictability, referring expressions, communicative efficiency.

*We are especially thankful to the authors who generously shared their datasets and provided clarifications via email, including those whose data we ultimately could not incorporate. Without their contributions, this research would not have been possible. We also thank Prof. Jennifer Arnold, Prof. Titus von der Malsburg, an anonymous reviewer, Associate Editor Prof. Morgan Sonderegger, and Editor Prof. John Beavers for their valuable feedback throughout the reviewing process. This research was supported by multiple funding sources. GB and TB received support from grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación. XL and LM received support from grant PID2021-122779NB-I00, funded by the Spanish Ministerio de Ciencia e Innovación, Agencia Estatal de Investigación/10.13039/501100011033 and the European Regional Development Fund—"A way of making Europe". TB was also supported by the Spanish Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación, and the European Social Fund Plus (ref. RYC2023-045215-I MCIU/AEI/10.13039/501100011033). XL, GB, TB and LM also received support from the Catalan government and the Department of Translation and Language Sciences at Universitat Pompeu Fabra (AGAUR grants 2020FI-B00575, SGR 2021 00470, SGR 2021 00947). XL was additionally supported by the Generalitat de Catalunya through the Aina project, and by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU –NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

1. INTRODUCTION. When people process language, they use contextual information to make predictions about what will come next. This predictive aspect of processing has received significant attention in recent research in Cognitive Science and Psycholinguistics (see a.o. Bubic et al. 2010; Clark 2013; Kuperberg & Jaeger 2016). The degree to which an addressee can anticipate what will come next based on contextual cues and prior knowledge is referred to as PREDICTABILITY. Models of language comprehension show that the more predictable some linguistic input is, the faster and more accurately it is processed by listeners (see, a.o., Smith & Levy 2013; Staub 2015; Kuperberg & Jaeger 2016).

As for language production, there is abundant evidence that speakers tend to use more reduced or attenuated forms for more predictable words or phrases (e.g., Jurafsky et al. 2001; Aylett & Turk 2004; Bell et al. 2009; Piantadosi et al. 2012; Jaeger & Buz 2017). Predictive processing frameworks suggest that these reductions enhance communicative efficiency by allowing for less speaker effort without incurring significant communicative cost. This in turn may also facilitate language processing for addressees (e.g., Levy & Jaeger 2006).

Given this background, it is natural to assume that predictability will influence the production of referring expressions: using pronouns is a more efficient way to refer to a previously mentioned entity that is already established in the discourse, especially when addressees can easily anticipate which entity the speaker is referring to.[1] However, the numerous studies that have explored this question in the last two decades have yielded mixed results. Some studies found that, indeed, speakers/writers are more likely to use pronouns when the referent is deemed to be more likely to be mentioned next (e.g. Arnold 2001; Rosa & Arnold 2017; Zerkle & Arnold 2019; Lindemann et al. 2020; Konuk & von Heusinger 2021; Weatherford & Arnold 2021; Hwang 2022; Medina Fetterman et al. 2022), while others did not find a significant difference in pronoun production between less predictable and more predictable referents (e.g. Ferretti et al. 2009; Fukumura & van Gompel 2010; Rohde & Kehler 2014; Holler & Suckow 2016; Mayol 2018; Kehler & Rohde 2019; Zhan et al. 2020; Frederiksen & Mayberry 2022; Hwang et al. 2022; Lam & Hwang 2022; Liao 2022; Patterson et al. 2022; Kravtchenko 2022; Hwang 2023). This discrepancy in findings has given rise to a long-standing debate in the field, as different theories make divergent predictions (see Section 2.1).

As a way to address this complex and inconclusive picture, we propose to synthesize the evidence obtained so far about this topic via a meta-analysis of 20 independent studies. The studies comprise 26 experiments, of which 8 found that predictability affects pronoun production and 14 did not find this effect (see Tables 1 and 6 in Section 3 for an

overview of these studies and their main findings). Through meta-analysis, we partially pool the results of the different studies, which enables us to estimate the overall effect size of predictability, and to examine the variability in effect sizes across studies. This kind of approach is particularly valuable when the literature is mixed or when studies have small sample sizes and low statistical power, as is the case here. Moreover, it can help identify potential moderators (i.e., variables that influence the strength or direction of the relationship) that may explain the heterogeneity in effect sizes across studies, such as task differences or characteristics of the stimuli.

A notable advantage of meta-analytic methods is the major increase in statistical power achieved by considering data from multiple studies jointly. In the present meta-analysis, data from over 1,100 unique participants were included. This starkly contrasts with any individual experiment on the topic, none of which exceeded 100 participants. This advantage bolsters the chances of detecting genuine effects.

Meta-analyses are commonly employed in disciplines that rely heavily on quantitative research, such as medicine (e.g., Xing et al. 2020; Casula et al. 2022), psychology (e.g., Anglim et al. 2020; Bockrath et al. 2022), economics (e.g., Heimberger 2020; Brown et al. 2024), and biology (e.g., Wu & Seebacher 2020; Wang et al. 2021). In contrast, their application in theoretical linguistics has been comparatively limited, primarily due to the field's smaller size and the traditionally restricted availability of experimental data. This is showcased in the analysis by Bochynsk and colleagues (2023), who randomly sampled 600 linguistics journal articles to assess the prevalence of transparency and reproducibility practices in linguistics research. Among the 600 articles, only four were identified as meta-analyses.

Although recent years have seen an increase of meta-analyses for language data, they have mostly appeared in applied or applied-adjacent fields within the language sciences, especially applied linguistics, speech language pathology, and language acquisition (see e.g. In'nami & Koizumi 2009; Yousefi & Biria 2018). There has also been a growing use of meta-analyses in cognitive science venues. Examples are Mahowald et al. 2016, examining the effect of syntactic priming in sentence production, Jäger et al. 2017, on the effect of retrieval interference in sentence comprehension, and Perret & Bonin 2019 or Bürki et al. 2023 on different aspects of picture naming. In contrast, meta-analyses remain rare in theoretical linguistics venues; a notable exception is Nicenboim et al. 2018 in the *Journal of Phonetics*. Together with Nicenboim and colleagues (2018), we argue that meta-analysis should be added to the standard toolkit of theoretical linguists such that it can inform debates about the theory of language.

In our meta-analysis, we find a small positive effect of referent predictability on

pronoun production, suggesting that speakers are slightly more likely to use pronouns for more predictable referents than for less predictable ones. In what follows, we present the background (Section 2), methods and results (Sections 3-4), and we include a thorough discussion (Section 5) where we explore the theoretical and methodological implications of our findings and propose directions for future research, before concluding (Section 6).

Data availability statement. All data processing and analysis code developed for this article is available at: `https://osf.io/qyahc`.

2. Background. Pronoun production is a complex process influenced by a a number of factors, and, as such, it has been the subject of extensive research. There is a wide consensus that pronoun production is affected by factors such as grammatical function (e.g., Crawley et al. 1990; Brennan 1995), information structure (e.g., Rohde & Kehler 2014), animacy of the referent (e.g., Fukumura & van Gompel 2011), recency of the mention (e.g., McCoy & Strube 1999; Ariel 2001), frequency of mention (e.g., Ariel 1990) and number of competitors in the previous context (e.g., Arnold & Griffin 2007).

While the effects of the aforementioned factors on pronoun production are uncontested, the effect of referent predictability is still under debate, despite much empirical and theoretical analysis. This has led to models of pronoun production that posit different kinds of relationships between referent predictability and pronoun use. Section 2.1 presents two prominent models of pronoun production and discusses how they make different predictions regarding the role of predictability. In Section 2.2, we survey a range of experimental manipulations that have been used to create contexts with varying levels of predictability. In Section 2.3, we discuss the scope and focus of our study.

2.1. Models of pronoun production and the role of predictability. This section presents two prominent models of pronoun production and interpretation: the Bayesian model (Kehler et al. 2008; Kehler & Rohde 2013) and the expectancy hypothesis (Arnold 1998, 2001, 2010).

The Bayesian model proposes that pronoun production and interpretation are not mirror images of each other, but that they are instead mediated by a third factor: the probability of the referent itself. This approach is able to capture some asymmetries that have been observed between pronoun production and interpretation since the pioneering work Stevenson et al. 1994. This research included some discourse completion experiments, in which participants had to read a sentence and write a continuation. When the discourse to be completed included a pronoun, like *He* in 1a, no clear bias was found as to which antecedent was implicitly picked; the pronoun was equally likely to refer to either the

subject or the indirect object antecedent in the previous sentence. In contrast, when participants were free to produce any referring expression they liked to continue the discourse, as in 1b, they overwhelmingly used pronouns to refer to the subject antecedent and proper nouns to refer to non-subject antecedents.

(1) a. Peter handed a book to Bob. He _____
b. Peter handed a book to Bob. _____

This situation seems puzzling at first sight: if speakers overwhelmingly produce pronouns to refer to subjects in 1b, why is it the case that when they need to interpret a pronoun in 1a they do not display a stronger subject bias? According to the Bayesian model, the reason is that interpretation is affected by how predictable the referent is regardless of the referring expression used to refer to it. Although pronoun production has a strong subject bias, the fact that the non-subject is highly predictable in 1 (i.e., it is very likely that participants will continue the story discussing what Bob did with the book; see Section 2.2 for extensive discussion on this kind of contexts) is able to affect pronoun interpretation, such that it is possible for the pronoun to refer to both antecedents.

The Bayesian model captures the relationship between interpretation and production through Bayes' Rule, as shown in Equation 1.

$$P(\text{referent} \mid \text{pronoun}) = \frac{P(\text{pronoun} \mid \text{referent})\, P(\text{referent})}{\sum_{\text{referent} \in \text{referents}} P(\text{pronoun} \mid \text{referent})\, P(\text{referent})} \tag{1}$$

The term on the left-hand side, $P(\text{referent} \mid \text{pronoun})$, is the pronoun interpretation bias, the probability that, given that a pronoun has been used, this pronoun refers to a particular referent. In contrast, the term $P(\text{pronoun} \mid \text{referent})$ represents the pronoun production bias: the probability that a pronoun will be used given that the speaker wants to refer to a particular referent. The crucial point is that these two probabilities do not mirror each other, but are mediated by a third probability, $P(\text{referent})$, the probability that a particular referent will be mentioned —that is, how predictable the referent is.

That this relationship holds between interpretation and production is known as the weak claim of the Bayesian model (Kehler & Rohde 2019). In contrast, the strong claim of the Bayesian model, henceforth STRONG BAYES (Kehler & Rohde 2013), takes the proposal one step further. Strong Bayes posits that referent predictability ($P(\text{referent})$), on the one hand, and pronoun production ($P(\text{pronoun} \mid \text{referent})$), on the other, are influenced by different contextual factors. Strong Bayes posits that referent predictability is primarily affected by semantic and pragmatic factors, such as verb types and coherence

relations. By contrast, factors determining pronoun production are grammatical and/or information structural, such as grammatical function or topichood. Consequently, strong Bayes predicts that a speaker's choice to pronominalize a referent will be unaffected by the set of semantic and pragmatic contextual factors that are known to influence referent predictability.

The claim of strong Bayes directly contrasts with the expectancy hypothesis (Arnold 2001), which proposes that both predictability and pronoun production are affected by semantic and pragmatic contextual factors. The expectancy hypothesis posits that referent predictability is closely tied to referent accessibility. In traditional approaches to discourse anaphora, accessibility (or salience) denotes the activation level of a referent in the interlocutors' mental representation of discourse and is thought to play a critical role in determining speakers' choice of referring expressions (e.g., Givón 1983; Gundel et al. 1993; Chafe 1994; Brennan 1995; Grosz et al. 1995). It has been commonly assumed that highly accessible referents are more often referred to using pronouns, whereas less accessible referents usually require more explicit expressions. Thus, by proposing a direct link between referent predictability and accessibility, the expectancy hypothesis predicts greater pronoun production for more predictable referents. More specifically, it suggests that the listener's estimate of the likelihood that a particular referent will be mentioned next strongly influences the activation level of that referent in the interlocutors' mental representation of the discourse. Thus, speakers can calculate the former as an estimate of the latter, choosing more reduced forms, such as pronouns, for referents that are expected or highly predictable to their listeners.

In the expectancy hypothesis, predictability can come from multiple sources, including both semantic and grammatical factors. Therefore, in this model the more frequent use of pronouns for subject referents can be related to the fact that referents in subject position tend to be highly predictable. However, it is uncontested that speakers prefer to use pronouns when the referent was last mentioned in subject position, and both the strong Bayes and expectancy hypothesis can correctly predict this behavior in natural discourse. Therefore, this grammatical effect does not help distinguish between the two models.

The divergence between the two instead becomes apparent in scenarios where the most predictable referent, given the semantics of the sentence, is not the one last mentioned in subject position (e.g., see Ex. 1, where the indirect object referent *Bob* is more likely to be mentioned next). In such cases, while the expectancy hypothesis predicts that grammatical factors still render subject referents highly predictable, it also posits that semantically-driven predictability will further influence pronoun production. Hence, a referent that is more predictable due to semantic reasons is more likely to be

pronominalized compared to one that is less semantically predictable, for both referents appearing as subjects and referents as objects in the previous context.

To sum up, strong Bayes predicts that semantically-driven predictability should not affect pronoun production, whereas the expectancy hypothesis predicts that it will.

2.2. Manipulation of referent predictability. To study the effect of predictability on pronoun production, previous studies have primarily employed story continuation tasks, with experimental items with varying degrees of predictability. As we will explain below in detail, the predictability of the referents is typically manipulated through semantico-pragmatic factors (verb type being the paradigmatic case). The advantage of this type of manipulation is that the predictability of the referent changes, while its structural properties are kept constant (in particular, the syntactic function and informational structural properties). These are the kind of cases in which strong Bayes and the expectancy hypothesis make diverging predictions: only according to the latter, but not to the former, should predictability influence production. In contrast, both agree that structural factors will have an effect on production, although the explanation of the mechanism why this is the case may vary in each proposal.

For these reasons, this meta-analysis focuses on the predictability that is primarily driven by semantic and pragmatic factors that play a role in establishing the coherence of the discourse. By doing so, we aim to provide a more precise and focused analysis of whether pronoun production is sensitive to semantically driven contextual biases that have been shown to influence expectations about the upcoming referent. This question lies at the heart of the difference between the expectancy hypothesis and the strong form of the Bayesian model, as outlined in the previous section.

In a typical experimental paradigm, participants are presented with a controlled context and asked to provide a natural continuation to it. As both comprehenders and producers, participants must first understand the context, such as Example 2 below from Arnold (2001), and then provide a continuation based on how they expect the story to proceed, as in Example 3. In these studies, referent predictability is operationalized as the frequency of a referent being mentioned as the grammatical subject in the matrix clause of continuations. Thus, the referent that is most frequently re-mentioned first in continuations is considered to be the more predictable one.

(2) There was so much food for Thanksgiving, we didn't even eat half of it. Everyone got to take some food home. Lisa gave the leftover pie to Brendan. _____

(3) Brendan loved pie and cakes and all manner of sweet things but didn't know how to bake.

While this methodology has faced criticism for being somewhat unnatural (Demberg et al. 2023; Ye & Arnold 2023), discourse completion tasks continue to be the predominant methodology used to investigate the impact of referent predictability on pronoun use. This is done by manipulating the context items so that they involve contrasting levels of referent predictability, while keeping their structural properties constant. Among these manipulations, varying the main verbs and discourse connectives has been most widely adopted in the literature (e.g., Arnold 2001; Fukumura & van Gompel 2010; Rohde & Kehler 2014; Rosa & Arnold 2017; Mayol 2018; Zerkle & Arnold 2019; Zhan et al. 2020; Hwang et al. 2022). Additionally, other less typical manipulation methods have also been used. For example, some studies manipulate the temporal structure of events (e.g., Ferretti et al. 2009), or the relative clause attached to direct objects (e.g., Kehler & Rohde 2019). We discuss them below.

VERB SEMANTICS. Referent predictability through verb semantics has focused on two types of verbs: transfer-of-possession verbs and implicit causality verbs (e.g., Arnold 2001; Fukumura & van Gompel 2010; Rohde & Kehler 2014; Mayol 2018; Zerkle & Arnold 2019; Weatherford & Arnold 2021).

Transfer-of-possession verbs express a transfer event and assign thematic roles of Source and Goal to participants in the event. The Source role identifies the object from which motion of transfer proceeds, while the Goal identifies the object towards which transfer proceeds (Stevenson et al. 1994). The reason why these verbs have been useful to check the role of predictability in pronoun production is that they are divided into two subgroups with symmetric argument structures: Source-Goal verbs such as *give* in 4, and Goal-Source verbs like *catch* in 5.

(4) Lisa$_{source}$ **gave** the leftover pie to Brendan$_{goal}$. _____

(5) Marguerite$_{goal}$ **caught** a cold from Eduardo$_{source}$ two days before Christmas. _____

Several story continuation experiments have shown a consistent tendency for participants to more frequently refer back to the Goal referent than to the Source referent, in both types of verbs (e.g., Stevenson et al. 1994; Arnold 2001; Rosa & Arnold 2017). According to Stevenson et al. 1994, this next-mention bias stems from a natural focus on the consequences of the previously described event. For instance, in 4, participants tend to

talk about the Goal referent *Brendan*, because the Goal is more salient than the Source when discussing the consequences of receiving the leftover pie. Building on Moens & Steedman 1988, Stevenson and colleagues (1994) suggest that the structure of an event consists of three components: the initial condition, the event itself, and the consequences. It is this last component that becomes the most highly focused in the representation of a transfer-of-possession event. The Goal becomes highly predictable because it is more prominent than the Source in the consequence component.

To test how this expectation bias towards the Goal over the Source influences pronoun production while controlling for the well-known effects of grammatical function, previous studies compare the pronominalization of the Goal and Source when both are introduced in the same grammatical position (e.g., *Lisa* in 4 vs. *Marguerite* in 5). It is precisely in this type of context where the expectancy hypothesis and strong Bayes make contrasting hypothesis: the former predicts that more predictable referents (*Marguerite* in 5) will be pronominalized more often than less predictable referents (*Lisa* in 4), while strong Bayes predicts that the pronominalization rate should be similar in both cases.

Implicit causality verbs, such as *impress* or *admire*, are another well-tested and frequently used verb type in manipulation. These verbs describe a mental state and assign two thematic roles: a Stimulus, which is the argument that gives rise to the psychological state, and an Experiencer, which is the argument that experiences the psychological state, as shown in 6.

(6) a. David$_{stimulus}$ **impressed** Linda$_{experiencer}$. _____
b. David$_{experiencer}$ **admired** Linda$_{stimulus}$. _____

Implicit causality verbs implicitly attribute the cause of the event to the Stimulus argument and studies have shown that these verbs increase the predictability of the Stimulus referent; that is, after one of these verbs, it is very likely that the Stimulus will be mentioned again to provide an explanation for the cause of an event (e.g., Stevenson et al. 1994; Fukumura & van Gompel 2010; Ferstl et al. 2011; Rohde & Kehler 2014; Mayol 2018; Zhan et al. 2020).

These verbs are particularly well-suited to study the effect of predictability because, like transfer-of-possession verbs, they present crossed argument structures: the Stimulus argument can be realized either as the subject, as in 6a, or as the object 6b. Thus, in the case of *impress* in 6a there is a strong preference for referring back to the Stimulus, *David*, in the subject position, while for *admire* in 6b, continuations also preferably refer to the Stimulus, *Linda*, which is in the object position. This has allowed researchers to test for the effect of predictability on pronominalization while controlling for the syntactic function

of the antecedent: are there more pronouns produced referring to the Stimulus referent *David* in 6a than to the Experiencer referent *David* in 6b? The expectancy hypothesis predicts a positive answer, since the Stimulus is more predictable than the Experiencer. In contrast, strong Bayes predicts that the rate of pronominalization should be similar in both cases.

Discourse relations. Researchers have used discourse relations as another factor to manipulate referent predictability in addition to verb semantics. Discourse relations hold between clauses and can be implicitly inferred or explicitly marked by connectives. For example, the statement *John left* can be connected to *Mary stayed* by Explanation (*John left (because) Mary stayed*) or Result (*John left (so) Mary stayed*). The biases we just discussed regarding transfer-of-possession verbs and implicit causality verbs are at play with the default discourse relations after these verbs: implicit causality verbs tend to elicit continuations with explanations and transfer-of-possesion verbs tend to elicit continuations with narrations. It is in those contexts in which the Stimulus and the Goal arguments, respectively, are highly predictable. However, previous research has shown that forcing another discourse relation is able to alter these default biases (e.g., Fukumura & van Gompel 2010; Holler & Suckow 2016; Hwang et al. 2022). For instance, while, as we have seen, speakers tend to continue segment 6a with the Stimulus, *Linda*, in an explanation (*because...*), they tend to continue it with the Experiencer, *David*, when talking about the result of the event (*so...*; see Example 7). This is because, when providing an explanation, it is very natural to talk about what the stimulus did to cause the event; in contrast, when discussing the results of the event, it is more likely that speakers will talk about the effects that the event had on the experiencer. Similar effects have been reported for transfer-of-possession verbs (Stevenson et al. 1994).

(7) David$_{experiencer}$ admired Linda$_{stimulus}$ **so** _____

While some studies focus on how discourse relations modulate transfer-of-possession and implicit causality biases (e.g., Stevenson et al. 1994; Fukumura & van Gompel 2010; Holler & Suckow 2016; Hwang et al. 2022; Hwang 2023), other studies extend the investigation of how discourse relations affect referent predictability to contexts beyond transfer-of-possession and implicit causality (Hwang 2022; Liao 2022). These studies explore the role of connectives in facilitating a general sense of subject continuity and action continuity. They found that connectives of Occasion/Narration, such as *and (then)*, better support this continuity than connectives of other relations, such as *while*, *as a result* or *but*. In other words, using *and (then)* to link clauses creates a stronger expectation for

a continuation of the same subject and action than using other types of connectives. For example, in Example 8 for a study about Korean by Hwang (2022), it was found that the subject was mentioned more often when the connective *and (then)* was used than when the connective *while* was used, indicating that the subject is more predictable in narrations than in other types of discourse relations.[2] Again, these are the type of contexts in which the predictions from the expectancy hypothesis and strong Bayes differ: the former would predict an increase of the pronominalization rate to refer to the subject in Narration as compared to Contrast; while the latter would predict no difference between the two contexts.

(8) Minswu-ka Hyenwu-wa palphyo cwunpi-lul **ha-ko/nuntey**
Minsu-NOM Hyunwoo-with presentation preparation-ACC do-and/while
_____
_____
'Minsu prepared a presentation with Hyunwoo and (then)'/ 'While Minsu was preparing a presentation with Hyunwoo_____'

In contrast to most of the previous studies employing story continuation tasks, Liao (2022) investigated the influence of discourse relations on referent predictability through the analysis of re-mention frequency in corpus texts. This method operates on the premise that hearers track statistical regularities in their input in order to predict upcoming information, and that corpus data capture the distributional patterns that have been used to give estimates of predictability (Frank et al. 2013; Verhagen et al. 2018; Guan & Arnold 2021). The contexts extracted by Liao (2022) from the corpus, while broader and more ecologically valid in and of themselves, contain more noise than the tightly controlled stimuli utilized in story continuation tasks.

Other manipulations of referent predictability. While manipulating either the verb or the connective is the most common strategy to achieve varying degrees of referent predictability, alternative approaches also exist. Since these approaches also manipulate predictability through semantic and pragmatic factors, we will include them in our meta-analysis, even if their approach is slightly less conventional.

One such method, used in Kehler & Rohde 2019, involves manipulating relative clauses attached to direct objects in object-biased implicit causality contexts (see Example 9). In their study, participants were found to produce fewer Explanation (e.g. those introduced by *because*) continuations for 9a than for 9b due to the relative clause in 9a already providing an explanation. As the object bias primarily arises within Explanation

continuations, the decrease in the number of Explanation continuations for 9a thus yielded a difference in next-mention biases, with fewer object re-mentions for 9a relative to 9b. In other words, *the patient* in 9b is more predictable than *the patient* in 9a. The authors attribute this outcome to the fact that most coherent relations that participants could use in these contexts, other than Explanation, tended to exhibit a stronger subject bias (Kehler et al. 2008).

(9) a. The doctor reproached the patient **who never takes her medicine**. _____

b. The doctor reproached the patient **who came in at 3pm**. _____

In another study, Ferretti and colleagues (2009) explored the effects of manipulating the temporal structure of events (see Example 10). They observed a higher propensity for participants to re-mention the Goal rather than the Source in Source-Goal contexts across both perfective and imperfective conditions, 10a and 10b. However, this bias towards the Goal was found to be reduced in 10b. In other words, the Goal *Bob* is more predictable in 10a compared to 10b. The authors attribute this finding to the higher salience of the Goal over the Source with respect to the end state of transfer-of-possession events (Stevenson et al. 1994; Arnold 2001), but the salience of the Goal is comparatively diminished in the imperfective condition, where the event is depicted as ongoing.

(10) a. John$_{source}$ **handed** a book to Bob$_{goal}$. _____
b. John$_{source}$ **was handing** a book to Bob$_{goal}$. _____

2.3. Scope of this study. Our study, following most of the previous studies, focuses on the subject of the target sentence: Analyzing which referent will be mentioned next in the subject position, and whether the subject is pronominalized. We do not consider pronominalization of other grammatical functions. As for referent predictability, while a range of factors can influence it, as explained above, this meta-analysis focuses on the predictability that is primarily driven by semantic and pragmatic factors that play a role in establishing the coherence of the discourse. As discussed above, these are the contexts that most clearly distinguish the predictions from the expectancy hypothesis and the strong form of the Bayesian model.

In contrast, the impact of grammatical and information structural factors (such as topichood) on the likelihood of re-mention and pronoun production is more firmly established (see, for instance, Centering Theory; Brennan et al. 1987; Brennan 1995). Empirical research within this domain mostly corroborates the widely recognized influence

of grammatical function on pronoun production, with more pronouns produced referring to subjects than non-subjects.

It is, however, worth noting that structural effects are less well-understood for the overt pronouns in null-subject languages. For example, while it has been reported that there is a clear division of labor between null and overt pronouns in Italian and Catalan (nulls having a strong preference for a subject antecedent and overts for a non-subject antecedent), the results are less clear in other languages, such as Spanish and Greek, in which the overt pronoun seems to be more flexible and exhibit less clear biases (Chamorro 2018; Mayol 2018; Torregrossa et al. 2020; Contemori & Di Domenico 2021). Our meta-analysis will include studies both from null-subject and non-null subject languages, as long as predictability is manipulated through meaning-driven factors.

In contrast, we have excluded studies that exclusively manipulated focushood in their experimental designs (Kaiser 2010) or employed manipulations that remain less well-defined in the literature, such as the information status and uniqueness status of referents (Brocher et al. 2018), order of mention (Fukumura & van Gompel 2015), frequency of referent nouns (Lau & Hwang 2016), indefiniteness by case-marking in Turkish (Özge et al. 2016), referent animacy (Fukumura & van Gompel 2011), referent specificity by pe-marking in Romanian (Chiriacescu & von Heusinger 2010), and social status of referents (Vogels 2019).

Lastly, our meta-analysis also excludes studies that operationalized predictability using information theoretic notions such as surprisal or entropy. These are studies in which human subjects or computational models are tasked to guess which referent will be mentioned next in a truncated corpus text (Tily & Piantadosi 2009; Modi et al. 2017; Aina et al. 2021). In these studies, referent predictability is a function of the proportion of subjects that correctly guess the upcoming referent; or of a language model's certainty about the correct upcoming referent. This line of work requires a different measure of the effect size of predictability than the majority of other studies, which use next-mention frequency as a way of quantifying referent predictability (see Section 3.2 for the calculation of effect size). While predictability is a dichotomous variable in story continuation tasks, it is a continuous one in referent guessing tasks. To make coefficients of logistic regressions with continuous and discrete-binary predictors comparable, one method is to standardize continuous predictor values by dividing them by two standard deviations (Gelman & Hill 2007). However, the necessary original datasets from Modi et al. 2017, including key contextual control variables such as distance between mentions or referent frequency, were not fully retained. Thus, we could not carry out the necessary standardization and reanalysis. Moreover, Aina et al. 2021 differs methodologically from the other studies by

relying exclusively on computational model predictions rather than human participants, making direct comparison less valid. While Tily & Piantadosi 2009 provided complete datasets, incorporating only a single study of this distinct methodological type would risk introducing confounding due to methodological heterogeneity. Thus, we decided not to include this study in our current analysis.

3. Method.

3.1. Study selection criteria. To present as comprehensive a picture of current research as possible, we conducted a literature search using a combination of keyword and forward methods (see Harari et al. 2020 for a summary of study identification methods). The full process is visualized in Figure 1.

As a first step (*Search* in Fig. 1), in January 2023, we located relevant studies in the academic search engine Google Scholar,[3] using the following two features: (i) they contain a combination of keywords: *refer* AND *pronoun* AND (*completion* OR *production*); (ii) they cite at least one of the four representative articles on this topic: Arnold 1998, Arnold 2001, Kehler et al. 2008, and Kehler & Rohde 2013.[4] This resulted in 776 unique articles for which we next conducted an abstract screening (*Screening* in Fig. 1).

[Figure 1 about here.]

We established the following criteria to include studies in our meta-analysis. First (criterion 1), the study uses a typical manipulation type in the field, with a focus on coherence-driven predictability (see Section 2.3 for study scope). Thus, the data collected in the study enables for the comparison of referring expression usage for more predictable referents and less predictable referents, while controlling for their grammatical function. Second (criterion 2), the candidate study codes both choice of next-mention and choice of referring expression. Third (criterion 3), the study investigates native, adult users' production of referring expressions.[5]

In the screening process, dissertations and conference papers were excluded if their analyses were also reported in a peer-reviewed article. In such instances, only the peer-reviewed article was considered for inclusion (e.g., Rohde & Kehler 2014; Weatherford & Arnold 2021). When encountering studies that lacked essential information for calculating effect sizes, we attempted to contact either the corresponding or first author to obtain unpublished data. Seven studies were ultimately excluded due to missing data or non-responsiveness from the authors. Consequently, 19 studies qualified for inclusion in the analysis. During our search, we became aware of one study that was not yet discoverable

online due to a delay in its publication. Despite this, we decided to include this study in our analysis.[6] Consequently, a total of 20 studies were incorporated in the final analysis.

Out of these 20 studies, 6 report multiple relevant experiments, each with independent samples, with slight variations in setting, or different stimuli. Specifically, Fukumura and van Gompel (2010) conducted an experiment where one group of participants freely chose the referent, while another group was instructed to continue with a specific referent. Weatherford and Arnold (2021) carried out two experiments that only differed in the order of character mentions in the context sentence. Similarly, Medina Fetterman and colleagues (2022) conducted two experiments, one in written format and the other in spoken format. Hwang (2023) manipulated predictability by varying connectives in one experiment and verb types in another. Solstad and Bott (2022) conducted two experiments using two different prompt types, connective prompts and full-stop prompts. Contemori and Di Domenico (2021) recruited both Italian and Spanish participants to complete the task in their own language.

To explore the influence of varying experimental conditions and materials on the effect size of predictability, we included these multiple experiments as separate samples in our analysis. As a result, 20 primary peer-reviewed studies comprising data from 26 samples (in total 1145 participants) were included in our meta-analysis, as summarized in Table 1 (see Table 5 in Appendix A for the publication type and sample size of each study).

[Table 1 about here.]

We also provide a brief summary of relevant studies we identified during the screening process, as well as relevant research work presented at conferences, along with their findings, to more clearly present the diverse and inconclusive picture in the current body of literature (for an overview, see Table 6 in Appendix A). Several studies have reported a positive effect of predictability on pronoun production. Among the studies on English speakers, Arnold 2001 and Weatherford & Arnold 2021 found this effect primarily for object referents, while Rosa & Arnold 2017, Zerkle & Arnold 2019, and one of the experiments in Ye & Arnold 2023 support this characterization more generally. Additional evidence comes from research on other languages, such as Korean (Hwang 2022), Spanish (Medina Fetterman et al. 2022, with effects only for overt pronouns), Romanian (Lindemann et al. 2020), and Turkish (Konuk & von Heusinger 2021, with effects only for subject referents).

In contrast, other studies have found no significant effect of predictability on pronoun use. This is the case for Ferretti et al. 2009, Fukumura & van Gompel 2010, Rohde & Kehler 2014, Kehler & Rohde 2019, Kravtchenko 2022, Liao et al. 2023, and the other

experiment in Ye & Arnold 2023. Cross-linguistic support for this characterization comes from studies on Catalan (Mayol 2018), Mandarin Chinese (Zhan et al. 2020; Hwang et al. 2022), German (Holler & Suckow 2016), Korean (Hwang 2023), and American Sign Language (Frederiksen & Mayberry 2022).

Additionally, some studies have indicated a more complex pattern. Lam & Hwang 2022 found an increased use of null pronouns for less predictable referents in Mandarin. Portele & Bader 2020, on the other hand, observed lower pronoun usage for both more predictable experiencer referents and less predictable stimulus referents, suggesting that predictability alone cannot fully explain these data.

3.2. Effect sizes: Odds ratios. In this meta-analysis, effect sizes are reported as *odds ratios*. The odds ratio is a statistical measure that evaluates the relationship between two properties in a population. It is frequently used when there are comparison pairs and when the variable of interest is dichotomous (for a more detailed introduction, see Sonderegger 2023). In our case, the comparison pairs are referents with higher predictability (e.g., the Stimulus referent *Alan* in *Paul liked Alan because ...*) contrasted with those with lower predictability (e.g., the Experiencer referent *Alan* in *Paul embarrassed Alan because ...*), while controlling for their grammatical function. The variable of interest is the use of pronouns versus other referential forms when re-mentioning these referents in subject position, which is dichotomous (e.g., Paul liked Alan because he/Alan ...).

To calculate an odds ratio, we first define the odds of an event occurring in each group and then take the ratio of these odds. An example of how odds ratios are calculated is given in Equation 2, on the basis of the illustratory fictitious data in Table 2. Specifically, the odds of using pronouns with more predictable referents is calculated as the ratio of the frequency of pronouns to the frequency of non-pronouns for this referent ($\frac{A}{C}$). Similarly, the odds for the less predictable referent is $\frac{B}{D}$, where A/B and C/D are the counts of pronouns and non-pronouns, respectively. The odds ratio is then calculated by taking the ratio of the two odds.

The resulting odds ratio of 2.67 indicates that pronouns are 2.67 times more likely to be used with the more predictable referent. An odds ratio of 1 would instead suggest no effect of predictability on pronoun production; and an odds ratio smaller than 1 suggests a negative effect of predictability.

[Table 2 about here.]

$$\text{odds ratio} = \frac{\text{odds for more predictable referent}}{\text{odds for less predictable referent}} = \frac{\frac{A}{C}}{\frac{B}{D}} = \frac{A \times D}{B \times C} = \frac{40 \times 20}{30 \times 10} = 2.67 \qquad (2)$$

In analogy to Table 2, for each experiment, we gathered the number of continuations and the pronominalization rate in each condition. In doing so, the same sample sometimes contributes multiple odds ratios. This happens under three different circumstances. First, when an experiment examined predictability while controlling for grammatical function of the antecedent. For instance, production data for minimal pairs constructed using Goal-Source verbs and Source-Goal verbs contribute two odds ratios: one that compares Goal-Subject to Source-Subject, and another that contrasts Goal-Object with Source-Object. The majority of studies included in our analysis exhibit this characteristic (Arnold 2001; Hwang et al. 2022; Hwang 2023; Medina Fetterman et al. 2022; Zerkle & Arnold 2019; Contemori & Di Domenico 2021; Fukumura & van Gompel 2010; Holler & Suckow 2016; Hwang 2023; Konuk & von Heusinger 2021; Mayol 2018; Patterson et al. 2022; Rohde & Kehler 2014; Solstad & Bott 2022; Weatherford & Arnold 2021; Zhan et al. 2020; Kehler & Rohde 2019; Portele & Bader 2020). Second, this also happens when a single sample is on null-subject languages, measuring production rates of both null and overt pronouns. This kind of study can be construed as contributing two effect sizes, one for overt, and one for null pronouns. An example is Medina Fetterman and colleagues (2022), who find that predictability primarily affects the use of Spanish overt pronouns but not of null pronouns. Other languages for which there is data for both null and overt pronouns are Catalan (Mayol 2018), Mandarin (Zhan et al. 2020; Hwang et al. 2022; Lam & Hwang 2022), Turkish (Konuk & von Heusinger 2021), Italian and Spanish (Contemori & Di Domenico 2021). Finally, a single sample may contribute multiple odds ratios when the experiment combines various types of stimuli or manipulates predictability in multiple ways (Holler & Suckow 2016; Hwang et al. 2022; experiments 1 & 2 in Hwang 2023; Experiment 1 in Solstad & Bott 2022). For instance, Hwang and colleagues (2022) conducted an experiment using both transfer-of-possession verbs (Source-Goal & Goal-Source) and implicit causality verbs (Experiencer-Stimulus & Stimulus-Experiencer). Analyzing the results within each verb category can generate at least one odds ratio, with the possibility of deriving additional odds ratios when also considering the previous two circumstances (e.g., four odds ratios for each verb category: production of null pronouns for subject referents, overt pronouns for subject referents, null pronouns for object referents, overt pronouns for object referents).

After teasing apart samples according to these three cases, we end up with a total of 102 effect sizes that speak to referent predictability. They draw from 20 peer-reviewed articles, comprised of 26 individual experiments/samples. The fact that some effect sizes draw from the same study or sample is reflected in the multi-level model structures we employ (see Section 3.3).

As highlighted by one of our reviewers, combining effect sizes from studies in a meta-analysis is not trivial. This presents several challenges, such as different dependent-variable structures (e.g., Morris et al. 2024). For instance, in studies on null-subject languages, the dependent variable *Referential Form* has three levels: overt pronoun, null pronoun, and other forms. In contrast, studies on non-null subject languages categorize the dependent variable into two levels: pronoun versus non-pronoun. In order to respect the structure of the dependent variable in null-subject language studies, one should use an analysis method like multinomial regression, where the dependent variable has three levels, and conduct separate meta-analyses for these studies (Riley et al. 2019). In this meta-analysis, we analyze studies on null-subject languages in two ways with a univariate dependent variable. We include two separate effect sizes: one for *Pronoun = overt* and one for *Pronoun = null*. Our primary interest is not the relative frequency of overt versus null pronouns, but whether null or overt pronouns, as opposed to other forms, are more frequently produced in one condition than another. This approach mirrors how the original studies analyze the data, treating them as if they were two separate experiments (e.g., performing one logistic regression for null pronoun vs. other forms and another for overt vs. other forms; e.g., Mayol 2018; Contemori & Di Domenico 2021). By doing so, we align more closely with the common dichotomous cut-off, allowing us to include these studies alongside non-null subject language studies in the meta-analysis to increase the statistical power. These studies, though different in their structure, fundamentally address the same research question. However, this approach may increase heterogeneity in the result (e.g., Chowdhury et al. 2020). To ensure that including studies on null-subject languages does not skew our findings later in Section 4, we conducted a sensitivity analysis excluding these studies, presented in Appendix H.1. We did not find evidence indicating that the inclusion of studies on null-subject languages qualitatively impact or distort our main results.

3.3. Analyses. Over the last decade, Bayesian meta-analysis has become increasingly prominent as an alternative to traditional frequentist methods. This approach offers advantages such as the integration of prior information and the ability to directly estimate the probability of hypotheses (see, for example, Nicenboim et al. 2018; Bürki et al. 2020; Thompson & Semma 2020). Building on this growing trend, all analyses in our study were performed using Bayesian inference methods, using the *brms*-package (Gelman et al. 2015; Bürkner 2021) in *R* (version 4.1.2, R Core Team 2021). All fits were run with four chains for 2000 iterations each, with half as warm-up. All fits were diagnosed to rule out pathologies in their estimates. All had parameters with a split $\hat{R} < 1.01$ (Vehtari

et al. 2021), suggesting well-mixed chains; they had no saturated trajectory lengths (i.e., the sampler did not stop prematurely); they had no divergent transitions (i.e., no difficulties in exploring the posterior); and they all had an energy Bayesian Fraction of Missing Information over 0.2 (i.e., no inefficiency in the momentum resampling between trajectories; Betancourt 2017).

Following Nicenboim and colleagues (2018), we conducted the meta-analysis in two stages (Higgins et al. 2019), as depicted in Figure 2. First, a within-studies stage assesses effect size and uncertainty of each individual study/comparison pair, arriving at a unique estimate of effect size in log odds per study/comparison pair. A second, between-studies, stage then infers a single grand overall effect size, based on the individual estimates (in log odds) from the first stage. The intuition behind this procedure is that the findings of each individual study are draws from a distribution of effect sizes particular to that study, with an (unobserved) true effect underlying it. This corresponds to stage one. Behind each true effect, particular to individual studies due to their own idiosyncrasies, however, meta-analyses assume an overall (again, unobserved) effect distribution from which they all draw. This corresponds to stage two. In this way, the goal of a meta-analysis is to estimate an overall effect of a subject matter based on the findings of individual studies; in our case, the overall effect of referent predictability on pronoun production.

[Figure 2 about here.]

More precisely, the first stage here consisted in estimating the effect suggested by each individual study with regard to pronoun use predicted by referent predictability using Bayesian logistic regression models. The models use weakly informative priors for both main effects/intercept ($\mu = 0$, and $\sigma = 1.5$) and random effects ($\mu = 0$, and $\sigma = 1$) (McElreath 2020). These priors were chosen to be weakly informative as a function of the scale of both the response and predictors (Lemoine 2019). Specifically, our response is binary (logit scale) and all predictors were coded as factors (see below).[7]

The binary predictor *referent predictability* was coded as 1 when a mention refers back to a more predictable referent (e.g., the stimulus referent *Alan* in *Paul liked Alan because Alan ...*), and as 0 for mentions of less predictable referents (e.g., the experiencer referent *Alan* in *Paul embarrassed Alan because Alan ...*). The dependent variable *pronoun use* was coded as 1 whenever a mention is realized using a pronoun (e.g., Paul liked Alan because he ...), and as 0 when other referential forms are used instead (e.g., Paul liked Alan because Alan ...). We analyzed the complete datasets from 19 experiments using Bayesian mixed effects models. In the case of the corpus study by Liao (2022), we incorporated random intercepts for document ID and verb type, mirroring the approach

of the authors. For the other 18 experiments, random intercepts were added for both participants and items, along with random slopes of the fixed effect by participant and item (Nicenboim et al. 2018). In Appendix B, we detail the sources we drew upon and whether the full dataset was available. For the 7 experiments where individual-level datasets were unavailable but mean pronoun production rates were reported, we fit Bayesian logistic regression models without random effects. In these cases, we generated synthetic data based on the reported mean pronoun production rates and sample sizes. Note that this approach cannot recover by-subject or by-item variability. This may lead to overconfident estimates and up-weight these studies in the meta-analysis. While we recognize this limitation, we chose this approach because the alternative—estimating by-subject and by-item variability from studies with available data and using these to simulate a more complex synthetic dataset for studies without individual-level data—would introduce additional assumptions about the homogeneity of subjects and items across studies. Given the variability in study designs, we believe this added layer of complexity would not necessarily improve our estimates and that it could, at worst, instead bias them in less transparent ways. Finally, we extracted posterior estimates of the effect (i.e., the estimated difference in pronominalization between referents that were more predictable and those that were less predictable).

As mentioned above, in the second stage, we estimate an overall effect based on the estimates from individual studies from the first stage. To do this, we performed multi-level meta-analyses with nested random effects that factor in the heterogeneity between studies. Because, as explained above, some articles reported results from multiple samples, population-level effects were included not only at the article level but also at the within-article level.[8] The models in this second stage use weakly regularizing priors. Specifically, we use Student-t's prior (3, 0, 2.5) for intercept and random effects, and flat priors for fixed effects so that we do not rule out extreme outcomes as much as with a normal.

We carry out two separate meta-analyses on separate datasets, with some overlapping data. The first analysis examines pronouns in non-null subject languages (such as German and English) and null pronouns in null-subject languages. The second analysis focuses exclusively on null-subject languages and includes data on both null pronouns and overt pronouns. The goal of the first one is to quantify the effect of predictability on the production of the most reduced referential form available in each language. This minimal form varies between Germanic languages, like English or German, where it is an overtly expressed pronoun, and null-subject languages, such as Catalan or Mandarin, where the most reduced form available is a null pronoun.

For this first meta-analysis, we constructed two separate models. The first model assumes that all included studies are comparable and that there are no important characteristics that distinguish them. In contrast, the second model, informed by the literature, identifies three potential moderators (i.e., variables that may influence the relationship between the independent variable *referent predictability* and the dependent variable *pronoun use*). These variables are:

1. Manipulation of referent predictability: implicit causality verbs, transfer-of-possession verbs, discourse relations, and relative clauses. Previous research has speculated that the influence of predictability may be context-dependent, with its effects being more pronounced in specific contexts, such as transfer-of-possession, while being more difficult to discern in others, such as implicit causality (e.g., Rosa & Arnold 2017).

2. Language family: Romance (Catalan, Italian, Spanish), Mandarin, Korean, Turkish, and Germanic (English, German). Pronouns exhibit varying behavior across different languages; for example, within null-subject languages, Mandarin and Romance null pronouns do not necessarily function in the same manner (e.g., Filiaci et al. 2014; Zhan et al. 2020). Consequently, the impact of predictability on pronoun usage may differ across language families.[9]

3. Grammatical function of the antecedent: subject or object. Some studies on English have found that the effect of predictability was stronger for references to the object than for references to the subject, arguing that this may be due to the overall high use of pronouns for subjects (e.g., Weatherford & Arnold 2021). However, when looking at Turkish, the pattern shifts. Konuk and von Heusinger (2021) observed that predictability affects the use of pronouns for referring to previous subjects, but this effect does not extend to objects. Therefore, we expect some interaction between Language family and Grammatical function of the antecedent.

Our second model factors in the influence of these three potential moderators by including them as predictors. In meta-analyses, these variables are termed *moderators* because they can be seen as moderating the dependent variable—here, the effect size of predictability—when included in a regression model. To facilitate interpretation of the results, all predictors were centered around their respective means. Specifically, we coded categorical variables using sum contrasts, such as the grammatical function of the antecedent, with *subject* coded as -1 and *object* coded as 1. Consequently, coefficients in the models represent the deviations of each level of a predictor from the mean across all levels.

We note that a recent study by Ye and Arnold (2023) also identified a difference in task modality (written vs. spoken tasks) in relation to referent predictability and pronoun usage. While the written task shows no effect of predictability on pronoun usage, the spoken task does, suggesting that the influence of predictability was more pronounced in communicative environments involving direct addressees. Task modality was not included as a moderator in our analyses due to data limitations. Specifically, the majority of available data were collected from written tasks, and the scarce spoken data were primarily in English (3 out of a total of 4 spoken studies). Moreover, our analysis also included a corpus-based study by Liao (2022), which used both written and spoken corpus data. Considering the data imbalance, accurately estimating model parameters would be challenging if we were to include *task modality* as a factor in our analysis. Therefore, we only performed an exploratory analysis of the impact of task modality on the relationship between referent predictability and pronominalization using a subset of the dataset (see Appendix I for the results and a more detailed discussion).[10] For English, there is also variability across the studies in terms of whether the analysis included zero pronouns (see Ex. 11 for an example of a zero pronoun from Zerkle & Arnold 2019). Zerkle and Arnold (2019) included zero pronouns in their dependent measure, while many of the other studies excluded them and only considered pronouns with phonological form (e.g., Arnold 2001; Rosa & Arnold 2017). Given that Zerkle and Arnold (2019) (also see Liao et al. 2024) indicated that analysis results do not vary much whether null responses are included or not, and considering the previously mentioned issue of data imbalance, we chose not to include this variable as a moderator in the current analysis and to leave it for future research.

(11) a. The duke received the basket from the duchess _____
b. and ∅ threw it down the hallway.

In the second meta-analysis, we focused specifically on null-subject languages. In our sample of studies, these are: Catalan, Italian, Korean, Mandarin, Spanish, and Turkish. These languages permit null subjects and consequently speakers can choose between two types of pronouns; null or overt. We consider them both in this analysis to assess potential differences in the impact of predictability. Indeed, previous studies have suggested that null and overt pronouns may be constrained by different factors and to varying degrees (Filiaci et al. 2014; Fedele & Kaiser 2015). For instance, Medina Fetterman and colleagues (2022) found that in Spanish predictability primarily affected overt pronoun production but not null pronoun production. Overt pronouns, though more explicit than null forms, still represent a reduced referential form compared to names or full noun

phrases. Null-subject languages thus provide a valuable context for investigating the consistency of predictability effects on the production of pronominal forms with varying degrees of reduction. This can contribute to our understanding of the semantic constraints governing the usage of diverse pronominal forms, a crucial aspect for developing reference production models in null-pronoun languages, as underscored by Medina Fetterman and colleagues (2022). Furthermore, given that previous research has largely focused on English, examining referential choices in languages with a broader range of referring expression types may reveal patterns that would otherwise remain undiscovered (Vogels 2019).

To assess variations in predictability effects between null and overt pronouns, we incorporated pronoun type as an additional predictor in this analysis. We also controlled for the three factors considered in the previous model: language family, grammatical function, and manipulation type. The individual effect sizes estimated during the first stage (within-study) can be found in Appendix E.

While publication bias is a common concern in meta-analyses, we did not expect it to play a major role in our study. This is because the question of whether predictability affects pronoun production is a debate with two sides. Studies might support the expectancy hypothesis (e.g., Arnold 2001), suggesting that predictability plays a role, or they might align with the strong Bayes perspective (e.g., Kehler & Rohde 2013), which suggests the opposite. Either outcome is equally valuable for the scientific discourse, thus both positive and negative results have a fair chance of being published. However, to be thorough and transparent, we still assessed the potential for publication bias using a funnel plot (Egger et al. 1997). We found no conclusive evidence of publication bias. We report our assessment of publication bias in Appendix D.

In the following results section, our primary focus will be on the outcome of the multi-level meta-analyses conducted during the second stage (between-study), as these synthesize the results from multiple studies and directly address our main research question.

4. Results.

4.1. Effect of predictability on the use of the most reduced reference form.

In the first analysis, we investigate how predictability influences the production of the most reduced referential form available in a language. This most reduced form corresponds to pronouns in non-null subject languages and null pronouns in null-subject languages, for instance.[11]

We begin with a population-level model with no individual-level predictors. That is,

this model only factors in variation across studies and within samples when estimating the overall effect of predictability across studies (26 independent experiments comprising 71 odds ratios; see Section 3.2 for details). This yields an overall estimated odds-ratio of 1.32 [1, 1.74, 95% CI], suggesting that the most reduced referential form is around 1.32 times more likely to be used for the more predictable referents than for the less predictable referents.[12] A Bayesian hypothesis test provides very strong evidence that the effect size of predictability is greater than 0, with a Bayes Factor (BF) of 35.3. This means the observed data are 35.3 times more likely under the hypothesis that the effect is positive than under the null hypothesis of no effect or a negative effect. By contrast, a BF smaller than 1 would indicate evidence favoring the null hypothesis.[13] Taken together, our findings suggest a positive effect of predictability on the production of the most reduced referential form. However, note that this is a very small to small effect, according to Cohen's guidelines (Cohen 1988) for qualitatively interpreting odds ratios (odds ratio $<$ 1.44 for very small effects and $1.44 \leq$ odds ratio < 2.48 for small effects).

The overall effect, together with the estimated odds ratios of the individual experiments, is shown in Figure 3. Note that these estimates do not correspond to those of the individual studies on their own, obtained in the first stage of our meta-analysis (See Figure 2). Instead, they are adjusted in light of the data from the other studies. This allows the estimates to draw strength from the data of other studies.

[Figure 3 about here.]

We also conducted a sensitivity analysis that excluded the corpus study (Liao 2022), reported in Appendix H.2. This is because this study used re-mention statistics from corpora, unlike the others, which are arguably noisier than laboratory-controlled stimuli, lacking strict control for features such as referent animacy. The results of the sensitivity analysis suggest that the inclusion of the corpus-based study introduces no bias to the meta-analysis.

4.2. Effect of predictability on the use of the most reduced reference form while controlling for potential variations. Recall that we study three variables that may explain variation across individual samples: grammatical function of the antecedent; manipulation type; and language family (see Section 3.3). We included these variables as fixed effects in our models and explored potential interactions among them.[14] To select the most suitable model, we used the leave-one-out cross-validation (LOO-CV; Vehtari et al. 2017), identifying the best model in terms of expected log predictive density (ELPD). That is, models are compared based on their expected

predictive accuracy on new data. Specifically, we started with a complex model, which included language family, grammatical function of the antecedent, manipulation type as fixed effects, as well as the interaction between language family and grammatical function of the antecedent, and the interaction between manipulation type and grammatical function of the antecedent.[15] We systematically reduced the number of predictors by removing one predictor at a time and assessed each removal by comparing the predictive accuracy and model parsimony with and without the predictor. The ranking of models in terms of ELPD is presented in Appendix F. We found that the best model was our initial one, which included all three variables as predictors, and interactions between language family and grammatical function of the antecedent, as well as between manipulation type and grammatical function of the antecedent.

Once these components were added to the model as predictors, the overall effect estimate increases to 1.36 [0.72, 2.59, 95% CI]. However, at the same time, the addition of these parameters to the model increases the uncertainty about the overall effect of predictability (note the wide 95% CI of [0.72, 2.59]). The overall effect, together with the estimated odds ratios of the individual experiments, is shown in Figure 4. Despite this uncertainty, the Bayesian hypothesis test continues to indicate evidence supporting a positive effect of predictability (BF = 6).

[Figure 4 about here.]

The summary of the model is given in Table 3. Our analysis does not provide clear evidence for any major influence of grammatical function of the antecedent, manipulation type or language family on the observed effect size. Nevertheless, the results suggest some trends indicating potential variation. Specifically, studies on Turkish tend to observe a larger effect of predictability compared to other language families (BF = 6), whereas studies on Mandarin indicate a tendency towards a smaller effect of predictability (BF = 7). In addition, the results suggest that the effect size of predictability on pronoun production tends to be smaller when the grammatical function of the antecedent was the object compared to when it was the subject (BF = 4). This difference was mostly driven by Turkish, where predictability has a particularly salient effect on references to the subject. On the other hand, in other languages, the effect of grammatical function of the antecedent is less conclusive or tends to manifest in the opposite way. This interaction between language family and grammatical function of the antecedent is visualized in Figure 5.

[Table 3 about here.]

[Figure 5 about here.]

4.3. Effect of predictability on the production of null pronouns and overt pronouns. This final analysis narrows down the meta-analysis to focus on null-subject languages only. The dataset for this focused analysis contains 69 odds ratios from 6 languages (Italian, Spanish, Mandarin, Korean, Turkish, and Catalan), obtained from 12 independent samples across 9 distinct studies.

As before, we include the three potential sources of variation as predictors (Section 3.3). One of the primary objectives of this final analysis was to examine the effect of predictability on overt pronouns and to compare it with the effect on null pronouns. Fitting a model without additional predictors does not provide much insight into this question, as it assumes that the effects on overt and null pronouns are the same. Therefore, we directly fit a model with moderators and add a new one called *pronoun type* that codifies the different pronominal forms (null and overt pronouns).

Again, we explored potential interactions among them. To select the most suitable model, we used LOO-CV, as described in the previous section. The full cross-validation results are presented in Table 13 in Appendix F. The best model included interactions between language family and grammatical function of the antecedent, between manipulation type and grammatical function of the antecedent, as well as between pronoun type and grammatical function of the antecedent.

The model summary is presented in Table 4.[16] Focusing on null-subject languages only, we obtain an effect estimate of 1.66 [0.73, 4.32, 95% CI]. This is larger yet more uncertain than the previous estimates that included non-null subject languages and excluded overt pronouns. Similar to previous results, the Bayesian hypothesis test continues to indicate support for a positive effect of predictability (BF = 9).[17] The forest plot displaying the overall effect and the estimated odds ratios of the experiments is shown in Figure 6.

[Figure 6 about here.]

Once more, we find no clear evidence supporting a major influence of grammatical function of the antecedent, manipulation type, or language family on the results. However, the results suggest that there are potentially some small variations in the effect size of predictability. First, predictability tends to have a smaller effect on referents with object antecedents in studies on Turkish ($\beta = -0.64$, 95% CI = $[-0.98, -0.30]$). Additionally, the effect of predictability on referents with object antecedents varies with the type of manipulation used in studies: it may be larger when predictability is manipulated using implicit causality verbs ($\beta = 0.26$, 95% CI = $[0.03, 0.48]$) and smaller when different discourse relations are used ($\beta = -0.24$, 95% CI = $[-0.58, 0.10]$). Second, there is a smaller effect of predictability on overt pronouns than on null pronouns ($\beta = -0.11$,

95% CI = $[-0.23, 0.02]$). There is also an interaction between grammatical function of the antecedent and pronoun type ($\beta = 0.15$, 95% CI = $[0.03, 0.27]$), suggesting that the effect on overt pronouns tends to be larger when the referent has an object antecedent. Third, strong evidence suggests that the effect of predictability tends to be smaller in implicit causality scenarios ($\beta = -0.41$, 95% CI = $[-0.90, 0.07]$), consistent with the speculations put forth by Rosa and Arnold (2017).

All in all, these findings are consistent with our previous results: the effect of predictability on pronoun production may be small, and the examined factors do not prominently explain the variation.

[Table 4 about here.]

One concern is the comparability of our Spanish data with data from other Romance languages included in the analyses (Italian and Catalan). Spanish data comes from two independent studies, Contemori & Di Domenico 2021 and Medina Fetterman et al. 2022. As noted in Table 1, the variety of Spanish tested in Contemori & Di Domenico 2021 may be considered a contact variety, and participants in Medina Fetterman et al. 2022 were from different countries/territories and had varying lengths of residence in the United States (see Table 5 in Appendix A). We performed accordingly a sensitivity analysis, excluding the Spanish experiments, to check the possibility that this introduces noise into the data due to possible variations in overt and null subject pronoun usage across different varieties of Spanish (e.g., Alfaraz 2015). The results indicate that the inclusion of Spanish data does not qualitatively impact or distort our results. For more details, see Appendix H.3.

5. Discussion. This article has addressed a debate within the existing literature that revolves around the influence of referent predictability (or next-mention expectation biases) induced by semantic and pragmatic factors on pronoun production. To address this question systematically, we conducted a meta-analysis synthesizing results from 20 independent studies, which comprise data from 26 experiments and 8 languages. Our primary objective was to investigate the effect of predictability on the production of the most reduced reference form, specifically pronouns in Germanic languages (English and German), as well as null pronouns in the case of null-subject languages (Catalan, Italian, Mandarin, Korean, Spanish, and Turkish).

Overall results. Our meta-analysis provides strong evidence in support of the hypothesis that referent predictability affects the use of the most reduced reference form, as opposed to the null hypothesis. However, the magnitude of this effect is very small

to small (Cohen 1988), which may explain the contradictory results in the literature. Specifically, the estimated overall effect is an odds ratio of 1.32 [1, 1.74, 95% CIs], indicating that the odds of using pronouns for more predictable referents are moderately higher than the odds of using pronouns for less predictable referents.

The estimated overall effect remains comparable in magnitude, although with heightened uncertainty (odds ratio of 1.36 with 95% CIs of [0.72, 2.59]), when accounting for potential sources of variation from grammatical function of the antecedent (subjects or objects), manipulation type (transfer-of-possession verbs, implicit causality verbs, discourse relations, and relative clauses), and language family (Germanic, Korean, Mandarin, Romance, Turkish). Although the 95% credible intervals straddle 1—which could suggest that the effect might not be present at all or go in the opposite direction—and indicate substantial uncertainty, a Bayesian hypothesis test indicates moderate evidence for a positive effect of predictability on pronoun production.

Our second analysis specifically targets null-subject languages, synthesizing data from 9 studies, comprising 12 experiments and 6 languages. This analysis factors in potential variation along the three variables mentioned above (grammatical function of the antecedent, manipulation type, and language family) as well as pronominal form (null vs. overt pronouns). This analysis continues to suggest evidence of a small overall effect of predictability (BF = 9), again with large uncertainty about the true effect (an odds ratio of 1.66 with 95% CIs of [0.73, 4.32]). In addition, we find evidence hinting at potential variations in the effect size arising from the grammatical function of the antecedent, language family, manipulation type, pronoun type, and their interactions. We discuss these findings in more detail below.

Taken together, the overall effects from our meta-analyses support the expectancy hypothesis (e.g., Arnold 2001, 2010), that is, the hypothesis that referent predictability does influence pronoun production. According to this framework, as the predictability of referents increases, the referents become more salient in the interlocutors' mental representation of discourse. As a result, speakers are more inclined to use more reduced forms for these referents, thereby signaling the addressees to retrieve the readily accessible referents from memory. Predictability, as posited by the expectancy hypothesis, can come from various sources including semantic, information structural, and grammatical constraints. The influence of predictability derived by structural and grammatical cues, such as grammatical function, on pronoun production is uncontested, with robust evidence for phenomena like the subject bias (e.g., Fukumura & van Gompel 2010; Rohde & Kehler 2014; Medina Fetterman et al. 2022). Our meta-analysis suggests that there is a (small) effect of semantically-derived predictability even when controlling for grammatical function

and information structural factors.

One possible explanation for our results is a trade-off in cognitive effort vs. communicative efficiency for speakers. If predictability estimates rely on semantic inference (Hartshorne et al. 2015) and are calculated dynamically, the process arguably imposes a substantial cognitive load on speakers, as they need to continuously assess and update the discourse model to determine the most predictable referent for their listeners. At the same time, semantic predictability as a cue offers communication efficiency and listener-oriented benefits. In line with the consistent but small effect that we find, speakers may utilize semantically-based predictability, but to a limited extent, balancing its costs and benefits during language production and comprehension.

Alternatively, predictability estimates may derive from statistical regularities in daily input and stored knowledge about referent re-mention frequencies, as discussed in some of the previous studies (e.g., Guan & Arnold 2021; Langlois et al. 2023; Johnson & Arnold 2023). For instance, if grammatical subjects are very frequently re-mentioned in general, speakers likely have this knowledge and expect listeners to anticipate the subject's re-mention. A similar reasoning could be applied to semantically-based predictability. In this case, relying on predictability is not supposed to impose much cognitive load on speakers. However, at present it is unclear if such predictability estimates are consistently recognized in specific scenarios like those involving transfer-of-possession or implicit causality verbs. For example, Liao and colleagues (2024) searched for corpus contexts that closely resembled the stimuli of transfer-of-possession and implicit causality verbs used in previous story continuation tasks, but found that they were infrequent in naturalistic language usage. This indicates that the predictability estimates in these scenarios might not be as readily internalized as grammatical predictability, and might still rely on semantic inference.

Limitations. While our meta-analysis suggests that there is evidence for a small positive effect of predictability on pronoun production, we should also stress that the case is by no means closed. The current meta-analysis, which is based on a sample of 20 studies, contains quite a bit of uncertainty in its estimates. Future work should enlarge its empirical basis as new studies are carried out.

It is also important to note that, in line with the evidence synthesized, we have made a conscious decision not to incorporate two potentially influential factors in this meta-analysis: character gender (same or different; Kravtchenko 2022; Medina Fetterman et al. 2022) and constraints on narrative continuation (whether participants were required to continue with one character or allowed the liberty to select the subsequent

mention; Bott et al. 2018; Kravtchenko 2022). The exclusion is based on the fact that including these variables would introduce additional levels to the analysis, and current available data is insufficient to accurately estimate the effect of these levels, leading to heightened uncertainty. We showed this in the two exploratory analyses we conducted; see Appendices J and K for the results and a more detailed discussion. Again, we encourage future research to explore these factors further when more data becomes available.

Recommendations for future research. To start addressing the heterogeneity present in the current research landscape, we examined the influence of four distinct variables on the relationship between predictability and pronominalization, namely grammatical function of the antecedent, language family, manipulation type, and pronominal type. While our study cannot conclusively attribute the divergent findings to any of this factors due to insufficient evidence, it identifies specific aspects where evidence is lacking, helping define the most promising paths to pursue.

First, it is important for future research to consider potential differences across language families and conduct more typologically diverse experiments. In our analysis, evidence of cross-linguistic differences primarily comes from Turkish in Konuk & von Heusinger 2021 (see Section 4.2), who reported an 86% null pronoun production rate for highly predictable subject antecedents, in contrast to a 49% rate for their less predictable counterparts. This 37% disparity between the two pronominalization rates in Turkish is noteworthy when compared to other language families, where the average difference is a mere 5.5%. However, this evidence comes from a single study, and it is necessary to conduct further experiments to assess cross-linguistic differences.

Second, we recommend that future research adopt experimental conditions that avoid eliciting pronominalization rates that fall near the bottom (approximately 0%) or the ceiling (approximately 100%), since such rates can limit the detection of the effects of referent predictability on pronoun production. For example, the production rate of null pronouns for object referents in experiments for languages such as Mandarin (Hwang et al. 2022) and Korean (Hwang 2023) is often close to 0. This near-zero production rate substantially limits the potential to observe any variations due to predictability: If null pronoun usage approaches the bottom, there is no room for further reduction due to predictability. Analogously, when pronominalization rates approach ceiling, it becomes challenging to observe any additional increase in pronoun usage as a result of predictability. This issue can be particularly pronounced in studies employing crowdsourced subjects since they are often paid per task, creating an incentive for them to complete tasks as quickly as possible. To optimize their task completion rate, workers may

favor the use of pronouns, which are less explicit but more time-efficient than other more time-consuming referring expressions. To address this issue, we recommend to conduct experiments in controlled lab settings or recruit participants from different sources. Future research could also explore alternative experimental conditions. For instance, most previous studies used narrative language in their experimental stimuli, which tends to elicit more frequent use of pronouns (especially to refer back to subject antecedents), potentially resulting in ceiling effects. Using alternative contexts, such as descriptive language, may yield a lower rate of pronoun use in general, providing a more varied use of referring expressions. Other methods adopted by previous research include selecting participants who exhibit variation of referring expressions in their production (see Rosa & Arnold 2017 for further discussion).

Future work is also necessary to explore which specific experimental conditions make the effect of predictability more or less likely to arise. In particular, it could be that the predictability effect is more salient in contexts involving transfer-of-possession verbs compared to implicit causality verbs, as hinted at by the weak evidence in our analysis (see Section 4.3). Rosa and Arnold (2017) highlighted several distinctions between the two verb types, as follows. First, implicit causality verbs like *admire* depict a feeling or a mental state and are considered atelic, meaning that they lack an inherent endpoint. In contrast, transfer-of-possession verbs like *give* are telic, entailing a clear endpoint. The presence of an endpoint in transfer-of-possession verbs may facilitate the conceptualization of events, making it cognitively simpler to understand and organize information related to them. This, in turn, could give rise to a more robust discourse model, characterized by a stronger and more stable mental representation during the process of language comprehension. Second, the coherence relations that support the goal effect in transfer-of-possession verbs are those that advance the narrative or outline the consequences of the initial event. Therefore, the chronological sequence of continuations mirrors the chronological order of events. In contrast, implicit causality effects primarily emerge in explanations and rely on pre-event information about the cause, which could be more difficult to access. Third, it is plausible that the experiencer, despite not being the implicit cause, holds particular prominence in the discourse. This prominence could be attributed to the fact that implicit causality verbs convey the experiencer's mental state. By emphasizing the experiencer's perspective, these verbs may inherently draw attention to the experiencer's role in the unfolding narrative. In light of all these distinctions, it is plausible for the predictability effect to be more salient in transfer-of-possession verbs, and future work should test this possibility and its implications.

Our study also highlights the need for an adequately large sample size. Recall that

the analysis suggests that the effect size of predictability is very small to small. The magnitude of the effect size is a key factor in determining the required sample size for an informative study. A smaller effect size is often associated with greater noise or variability in the data, thereby making it more challenging to detect a difference. For studies on the role of predictability, thus, a larger sample size than is usual in the field is needed. Additionally, the small effect size of predictability implies that its influence might be overshadowed by more potent factors, such as the well-established effects of grammatical functions, as noted by Vogels (2019). This highlights the importance of controlling for these factors in the experiments.

Another very important avenue for future research is the role of individual differences in the predictability effect. If tracking predictability imposes an increased cognitive burden on speakers, those with greater cognitive resources, such as higher working memory capacity, may be better at tracking referent predictability and using this information to plan and produce upcoming utterances. Assessing the role of individual differences, in addition to being intrinsically relevant as part of language variation, is useful to distinguish between alternative explanations of speakers' behavioral patterns. Our discussion above offers two alternative hypotheses (trade-off between cognitive and communicative pressures vs. use of statistical patterns). A link between cognitive resources and linguistic behavior would favor the first.

Finally, we strongly advocate for adopting open science practices in linguistics, in particular the release of experimental data together with the analysis. These practices enhance transparency, credibility, and replicability of research findings, and notably facilitate collaboration. By making data accessible, researchers enable others to build upon existing work, combine datasets, and conduct more powerful analyses, as demonstrated by the present meta-analysis. A vantage point provided by meta-analytic approaches is showcased in Figure 3. While no individual study shows a 95% CI that excludes 1, the overall effect estimated by our meta-analytical model provides rather strong evidence for a positive effect of predictability. In other words, no single study included in this meta-analysis offers conclusive evidence for a predictability effect—despite many reporting such results individually. However, when combined, the studies suggest a stronger signal, demonstrating the potential value of meta-analysis in synthesizing fragmented evidence. By making data accessible through open science practices, the field can enable more comprehensive and reliable conclusions like this.

6. Conclusion. Our study, being the first meta-analysis on the topic of referent predictability and pronoun usage, carries both methodological and theoretical implications

for research on this topic and linguistics studies more broadly. By integrating and synthesizing findings across various independent studies, meta-analyses offer a powerful tool for identifying overarching patterns and addressing inconsistencies in the literature. Tacking stock, our synthesis of the present empirical landscape suggests that the effect of coherence-driven predictability on pronoun production is likely positive and small. This implies that speakers use predictability as a cue, which listeners are sensitive to, at least to some degree. Additional research is required to explore potential variations in the observed effect under diverse conditions and across diverse contexts. In broader terms, we hope that this study can also serve as an example for future studies to build on. The data used in this analysis (openly available) will enable future researchers to pose and address questions of their own design, fostering continued meta-analysis and data aggregation in the areas of predictability and pronoun production. The methodology (as well as the analysis software, which we also release) can be adapted to other areas of study in linguistics. We hope that our study will help making meta-analysis a standard tool in linguistics, which would enable a better connection between data and theory.

Appendix A More information about the relevant studies. Table 5 provides information about the studies included in our meta-analysis, complementing the data in Table 1.

[Table 5 about here.]

Table 6 presents information about the relevant studies that address our research question. Some studies are not included in Table 1 and Table 5 due to missing data, which prevents us from estimating the effect size for our meta-analysis. On the other hand, Contemori & Di Domenico 2021 and Solstad & Bott 2022 are listed in Table 1 and 5 but not in Table 6 because they do not directly address our research question.

[Table 6 about here.]

Appendix B Dataset availability. In our meta-analysis, we included a total of 26 experiments. Among these, 19 experiments provide complete datasets, allowing us to fit Bayesian mixed logistic models with random intercepts added for participants and items, as explained in Section 3.3. For the remaining 7 experiments, where we only have mean pronoun production rates, we fit Bayesian logistic models without additional random structures. The dataset availability is summarized in Table 7.

[Table 7 about here.]

Appendix C Prior sensitivity analysis. To assess the influence of prior choice on our models, we compared the outcomes using weakly informative priors against those using *brms* default priors. Our selected priors for both main effects and intercept were $\mu = 0$ and $\sigma = 1.5$, and for random effects, $\mu = 0$ and $\sigma = 1$. The *brms* default priors used flat priors for fixed effects and Student's-t priors (df = 3, $\mu = 0$, $\sigma = 2.5$) for intercepts and random effects.

Our priors were chosen based on the literature: the effect of predictability, if present, is likely to fall within the range of [-4.5, 4.5] in terms of the odds ratio. Cases where the odds of pronoun usage for more predictable referents are 350% higher than those for less predictable referents are very expected to be rare. This is especially true for subject referents, where pronoun production is often already high, leaving limited room for a substantial increase in pronoun usage for more predictable referents compared to less predictable ones. In this way, our priors are informative but weak; in the sense of not constraining the effect too much and allowing for it to be both negative and positive. The alternative offered by the default flat priors of *brms* could unnecessarily reduce statistical

power, allocating mass to extreme values that are implausible given the mixed evidence from the field. If the effect of predictably was very large then we would expect less of a debate to exist.

Following the guidance in Chapter 3.6 of Nicenboim et al. 2024, we conducted prior predictive checks. Specifically, we fit each of our *brms* models with `sample_prior = "only"`, instructing the model to sample exclusively from the prior distributions without incorporating the data. We then visually assess whether the priors are reasonable. As an example, we illustrate this process using the study by Rohde and Kehler (2014). Figure 7 depicts observed data with a black thick line y and the (prior-)predicted data with light blue lines y_{rep}, showing 100 samples from the prior predictive. The priors appear to be reasonable. While the simulated y_{rep} values broadly align with the observed data, they also allow for ample room for the posterior to be informed by the data. This qualitative pattern of the priors is consistent across all studies, with the exception of two models that we excluded due to insufficient data points (see below). For prior predictive check plots of the remaining studies, please see our OSF repository.

[Figure 7 about here.]

When comparing our priors and the *brms* default priors, we observed no substantial differences in the parameter estimates. However, models using *brms* default priors showed wider 95% credible intervals for fixed effects, indicating increased uncertainty. For illustration, we refitted Bayesian logistic regression models for the study by Rohde and Kehler (2014) with both priors, presented in Table 8.

[Table 8 about here.]

Our chosen priors also helped to identify datasets inadequate for reliable analysis. For instance, models for null pronoun production for object referents in Mandarin (Hwang et al. 2022) and Korean (Hwang 2023) failed to converge using our priors due to insufficient data points, whereas the same models using *brms* default priors converged but yielded highly uncertain estimates, as shown in Table 9. Therefore, we excluded these datasets from the meta-analysis.

[Table 9 about here.]

Appendix D Publication biases. We assessed the potential for publication bias visually with a funnel plot, shown in Figure 8, generated by the *funnel* function from the *meta* package (version 5.2.0; Balduzzi et al. 2019).

[Figure 8 about here.]

We further conducted Egger's test (Egger et al. 1997) to test the funnel plot asymmetry using the *metabias* function from the same package. The results suggest a potential minor publication bias toward reporting positive effects (β = -0.53, p = 0.07). However, the asymmetry is not strong enough to provide conclusive evidence of publication bias.

Appendix E Estimated effect size from individual comparison pairs. In the first stage of our meta-analysis, we computed summary statistics for each comparison pair (more predictable referents vs. less predictable referents). We did this separately for subject referents (e.g., *Paul* in *Paul embarrassed Alan because Paul ...* vs. *Paul* in *Paul liked Alan because Paul ...*) and object referents (e.g., *Alan* in *Paul liked Alan because Alan ...* vs. *Alan* in *Paul embarrassed Alan because Alan ...*) to control for the well-established effects of grammatical functions on pronoun production.

The effect of each individual comparison pair was estimated using Bayesian (mixed) logistic models, in which pronoun use was predicted by referent predictability (binary factor; whether a referent is more or less predictable in a pair), as explained in Section 3.3. For null-subject languages, we constructed separate models for overt pronouns (use of overt pronouns predicted by whether a referent is more or less predictable in a pair) and null pronouns (use of null pronouns predicted by whether a referent is more or less predictable in a comparison pair). We extracted posterior draws from the regression models and presented the means and 95% credible intervals in Table 10 for subject referents, and Table 11 for object referents.

[Table 10 about here.]

[Table 11 about here.]

Appendix F Model selection: Leave-One-Out Cross-Validation. In Section 4.2, we described how the model selection was carried out using leave-one-out cross-validation (LOO-CV). This section details the LOO-CV results and the difference in expected log predictive density (ELPD) between each model and the best model.

The ranking of the models is shown in Table 12. The best model, shown in the top row as *all predictors*, includes language family, grammatical function of the antecedent, and manipulation type as fixed effects, with interactions between language family and grammatical function of the antecedent, and between manipulation type and grammatical function of the antecedent included. A more complex model, which included a fuller set of

interaction terms (specifically, language family, grammatical function of the antecedent, and manipulation type as fixed effects, along with their pairwise interactions), failed to converge.

[Table 12 about here.]

Similarly, to select the best model for null-subject languages, we employed the same LOO-CV approach, as described in Section 4.3. The LOO-CV results are presented in Table 13. The best model includes grammatical function of the antecedent, pronoun type, manipulation type, language family as fixed effects, as well as interactions between pronoun type and grammatical function of the antecedent, between grammatical function of the antecedent and manipulation type, and between language family and grammatical function of the antecedent.

[Table 13 about here.]

In this selected model, a potential concern is the relatively large number of fixed-effect terms compared to the number of data points (see Appendix G for details). Such a configuration can increase the risk of overfitting. To address this, we conducted sensitivity analyses by systematically reducing the number of fixed-effect terms and comparing the resulting estimates. As shown in Table 14, the estimated overall effect consistently supports a positive effect of predictability across models. This consistency suggests that our conclusions are robust to the inclusion or exclusion of specific terms.

[Table 14 about here.]

Appendix G Number of datapoints by each predictor and its levels.

Table 15 displays the number of datapoints categorized by predictor and level analyzed in Section 4.2, which examines the effect of predictability on the use of the most reduced reference form.

Table 16 reports the number of datapoints categorized by predictor and level examined in Section 4.3, focusing on the effect of predictability on the production of null pronouns and overt pronouns.

[Table 15 about here.]

[Table 16 about here.]

Table 17 shows the data structure of the dataset analyzed in Section 4.2. The sparsity of certain combinations of language family and manipulation type likely contributed to the non-convergence of the full model, which included an interaction between these two predictors.

[Table 17 about here.]

Appendix H Sensitivity analyses.

H.1 Excluding studies on null-subject languages. We exclude studies on null-subject languages which has different dependent-variable structures (three levels) from studies on non-null-subject languages (two levels) to ensure that including studies on null-subject languages using our approach does not skew our findings in Section 4.1 and 4.2.

We refit a basic random-effects model without covariates for the remaining 11 studies on Germanic languages (14 independent experiments). By analyzing the effect of predictability in this manner, we obtained an overall estimated odds ratio of 1.31 [1, 1.70, 95% CIs], indicating that the most reduced referential form is 1.31 times more likely to be used for more predictable referents compared to less predictable referents. This effect size is very close to the one reported in Sections 4.1 (1.34 [1, 1.75, 95% CIs]), where we included the studies on null-subject languages.

Next, we refitted the model with the studies on null-subject languages excluded and included (a) grammatical function of the antecedent, (b) manipulation type and the interaction between them as covariates to account for potential sources of variation. Language Family was not added because the remaining studies were all studies on Germanic languages. After adjusting for these factors, the effect estimate was 1.26 [0.85, 1.82, 95% CIs], remaining small while exhibiting greater uncertainty. Bayes Factors indicate strong evidence in favor of a positive effect of predictability (BF = 9). The summary of this model is presented in Table 18. With the inclusion of studies on null-subject languages, there remains no clear evidence of any major influence from the added covariates.

[Table 18 about here.]

In conclusion, the sensitivity analysis demonstrates that the inclusion of the studies on null-subject languages does not introduce major biases into our results.

H.2 Excluding corpus study. In the subsequent analyses, we exclude the corpus study Liao 2022 and focus only on the data obtained in psycholinguistic experiments to rule out the possibility that including the corpus study biases the results or drives the uncertainty.

We refit a basic random-effects model without covariates for the remaining 25 independent experiments (comprising 70 odds ratios). By analyzing the effect of predictability in this manner, we obtained an overall estimated odds ratio of 1.35 [1, 1.80, 95% CIs], indicating that the most reduced referential form is 1.35 times more likely to be used for more predictable referents compared to less predictable referents. This effect size is very close to the one reported in Sections 4.1 (1.32 [1, 1.74, 95% CIs]), where we included the corpus study.

In conclusion, the sensitivity analysis demonstrates that the inclusion of the corpus study Liao 2022 does not introduce biases into our results.

H.3 Excluding Spanish studies. In this section, we address concerns regarding the comparability of Spanish data from Contemori & Di Domenico 2021 and Medina Fetterman et al. 2022 to data from other Romance languages in our analyses (see main text for an explanation of these concerns). We exclude the Spanish data and reanalyze the dataset to examine the potential impact on our results and the associated uncertainty. To estimate the effect of predictability on the use of the most reduced form, we begin by refitting a basic random-effects model without covariates for the remaining 23 independent experiments (comprising 61 comparison pairs). This approach yields a pooled estimated odds ratio of 1.38 [0.98, 1.92, 95% CIs], which is again very similar to the estimate reported in Sections 4.1 (1.33 [1, 1.75, 95% CIs],) when including the Spanish data.

Next, we refit the model with the Spanish data excluded and incorporate (a) grammatical function of the antecedent, (b) manipulation type, (c) language family, and (d) interactions between language family and grammatical function of the antecedent as well as between manipulation type and grammatical function of the antecedent as covariates to account for potential sources of variation. After adjusting for these factors, the effect estimate increases to 1.54 [0.76, 3.22, 95% CIs], indicating greater uncertainty. However, Bayes Factors continue to indicate evidence supporting a positive effect of predictability (BF = 8). The summary of this model, presented in Table 19, displays similar estimates to those in Table 3 when the Spanish data was included in the analysis. Also, there is still no strong evidence supporting the influence of grammatical function of the antecedent, manipulation type, or language family on the results. Besides, the results

continue suggesting that studies on the Turkish language tend to observe a larger effect of predictability on references to the subject.

[Table 19 about here.]

Finally, we refit the model specifically for null-subject languages without the Spanish data. The overall effect is estimated to be 1.52 [0.44, 5.37, 95% CIs], similar in magnitude but with much greater uncertainty than the original model reported in the main text (Table 3). With the reduced data, Bayes Factors continue to support a positive effect of predictability, albeit with weaker evidence than before (BF = 4). The summary of the model is provided in Table 20. Although the estimates are similar to those in the original model, the estimate for the Romance languages exhibits larger uncertainty due to the reduced sample size.

[Table 20 about here.]

Overall, the sensitivity analysis demonstrates that the inclusion of Spanish data from Contemori & Di Domenico 2021 and Medina Fetterman et al. 2022 does not particularly impact or distort our results.

H.4 Excluding two studies that aimed at addressing other related questions. Contemori and Di Domenico (2021) and Solstad and Bott (2022) did not conduct experiments specifically aimed at addressing our research question. However, we included them in our analysis as they employed the same experimental paradigm as other studies and their data allowed for the calculation of the effect size of predictability. Here we exclude their data and reanalyze the dataset to examine the potential impact on our results and the associated uncertainty.

To estimate the effect of predictability on the use of the most reduced form, we begin by refitting a basic random-effects model without covariates for the remaining 22 independent experiments (comprising 61 comparison pairs). This approach yields a pooled estimated odds ratio of 1.30 [0.94, 1.79, 95% CIs], which is very close to the estimate reported in Sections 4.1 (1.33 [1, 1.75, 95% CIs]) when including the data from these two studies.

Next, we refit the model with these two studies excluded and incorporate (a) grammatical function of the antecedent, (b) manipulation type, (c) language family, and (d) interactions between language family and grammatical function of the antecedent as well as between manipulation type and grammatical function of the antecedent as covariates to account for potential sources of variation. After adjusting for these factors,

the effect estimate increases to 1.38 [0.70, 2.77, 95% CIs], indicating greater uncertainty. As before, Bayes Factors continue to indicate support for a positive effect of predictability (BF = 5).

The summary of this model, presented in Table 21, displays similar estimates to those in Table 3 when these two studies were included in the analysis. Also, there is still no strong evidence supporting the influence of grammatical function of the antecedent, manipulation type, or language family on the results.

[Table 21 about here.]

Finally, we refit the model specifically for null-subject languages without these two studies. The overall effect is estimated to be 1.62 [0.55, 5.00, 95% CIs], with great uncertainty. As before, Bayes Factors continue to indicate support for a positive effect of predictability (BF = 6). The model summary is provided in Table 22. The estimates are very similar to those in Table 3.

[Table 22 about here.]

Overall, the sensitivity analyses demonstrate that the inclusion of data from Contemori & Di Domenico 2021 and Solstad & Bott 2022 does not particularly impact or distort our results.

H.5 Using individual languages as a covariate instead of language families. In this section, we present a sensitivity analysis examining the impact of using individual languages as opposed to language families as a covariate. When incorporating language family, manipulation type, grammatical function of the antecedent and their interactions as covariates, the summary estimate of the effect of predictability on the most reduced reference form is 1.37 [0.76, 2.51]. When using individual languages instead of language families as a covariate, the estimate increases to 1.40 [0.72, 2.8], accompanied by a slightly higher degree of uncertainty (see Table 23).

Furthermore, the effect of predictability in null-subject languages is estimated to be 1.54 [0.50, 4.62], when individual languages are utilized as a covariate instead of language families (refer to Table 24), which also exhibits much increased uncertainty. Estimates on individual languages for which only a single study is available are particularly uncertain (e.g. for Catalan, Italian). Despite this, Bayes Factors continue to indicate support for a positive effect of predictability (BF = 5).

In conclusion, utilizing individual languages rather than language families as a covariate in the analysis leads to increased uncertainty in the results, but does not particularly alter our results.

[Table 23 about here.]

[Table 24 about here.]

Appendix I Exploratory analysis: the influence of task modality on the relationship between referent predictability and pronoun use. In a recent study by Ye and Arnold (2023), the impact of predictability on pronoun usage was found to be dependent on task modality. The study reported an influence of implicit causality in spoken tasks, but not in written tasks, suggesting that the effect of predictability is more pronounced in interactive communicative contexts.

We conducted an exploratory analysis with task modality (written vs. spoken) as the sole covariate, restricting our examination to English and Spanish, the languages for which both spoken and written experimental data were available (8 studies, 11 experiments). Task modality was centered to ensure that the coefficients in the model represented deviations from the mean for each level.

By incorporating modality as a main-effect variable, the overall effect of predictability was estimated to be 1.27 [0.91, 1.77, 95% CIs]. The resulting model summary is presented in Table 25. Our analysis does not offer clear evidence for any significant influence of modality. However, the Bayesian hypothesis testing does suggest that there is a high probability (85%) that the influence of predictability on the most reduced form might tend to be less pronounced in written experiments compared to spoken ones. Specifically, the data are approximately 5.83 times more likely under the hypothesis that the effect size of predictability is smaller in the written modality compared to the alternative hypothesis. This corroborates the findings by Ye and Arnold (2023). However, the available evidence in our analysis is insufficient to draw definitive conclusions.

[Table 25 about here.]

Appendix J Exploratory analysis: the influence of character gender on the relationship between referent predictability and pronoun use. Kravtchenko (2022) argues that an effect of predictability on production is more likely to be found using same-gender prompts than using different-gender prompts and it would be expected if speakers are choosing referring expressions based on expected communicative utility (see more detailed discussion in Kravtchenko 2022).

We conducted an exploratory analysis with character gender (same vs. different) as the sole covariate, restricting our examination to non-corpus studies as the character gender in corpus texts was not manipulated. In total, we have 24 datapoints for different-gender

conditions, and 46 for same-gender conditions. Gender was centered to ensure that the coefficients in the model represented deviations from the mean for each level.

By incorporating character gender as a main-effect variable, the overall effect of predictability was estimated to be 1.32 [0.95, 1.82, 95% CIs]. The resulting model summary is presented in Table 26. Our analysis does not offer clear evidence for any significant influence of gender. However, the Bayesian hypothesis testing suggests that there is a high probability (84%) that the influence of predictability on the most reduced form might tend to be less pronounced in the different-gender condition compared to the same-gender one. Specifically, the data are approximately 5.35 times more likely under the hypothesis that the effect size of predictability is smaller in the different-gender condition compared to the alternative hypothesis. This effect is in the predicted direction by Kravtchenko (2022). However, the estimate -0.09 also indicates that the difference, if any, might be very small, and the available evidence in our analysis is insufficient to draw definitive conclusions.

[Table 26 about here.]

Appendix K Exploratory analysis: the influence of constraints on narrative continuation on the relationship between referent predictability and pronoun use.

It has been hypothesized that the effect of predictability of pronominalization may be affected by the type of completion paradigm in the experimental task, which can be free or constrained. In a constrained completion paradigm, it is indicated to participants which referent should be mentioned first in the continuation. In free completion, participants are free to complete the passage as they wish, mentioning either one of the antecedents (or something else entirely) first.

Bott and colleagues (2018) argue that if one assumes that participants avoid lower-predictability referents in free-completion tasks, but are forced to refer to them in constrained-completion tasks, then effects of predictability on pronominalization should arise given constrained completion paradigms, and not (or to a lesser degree) in free completion paradigms.

We conducted an exploratory analysis with completion paradigm (constrained vs. free) as the sole covariate. In total, we have 42 data points for the free-completion paradigm, and 29 for the constrained one. Completion paradigm was centered to ensure that the coefficients in the model represented deviations from the mean for each level.

By incorporating completion paradigm as a main-effect variable, the overall effect of predictability was estimated to be 1.36 [1, 1.86, 95% CIs]. The resulting model summary is presented in Table 27. Our analysis does not offer clear evidence for any major influence

of free or constrained completion paradigm. However, the Bayesian hypothesis testing suggests that there is a high probability (82%) that the influence of predictability on the most reduced form might tend to be less pronounced in the free-completion condition compared to the constrained one. Specifically, the data are approximately 4.48 times more likely under the hypothesis that the effect size of predictability is smaller in the free-completion condition compared to the alternative hypothesis. This effect is again in the predicted direction. However, the available evidence in our analysis is insufficient to draw definitive conclusions.

[Table 27 about here.]

REFERENCES

Aina, Laura; Xixian Liao; Gemma Boleda; and Matthijs Westera. 2021. Does referent predictability affect the choice of referential form? A computational approach using masked coreference resolution. *Proceedings of the 25th Conference on Computational Natural Language Learning*, 454–469. Online: Association for Computational Linguistics. Online: https://aclanthology.org/2021.conll-1.36.

Alfaraz, Gabriela G. 2015. Variation of overt and null subject pronouns in the Spanish of Santo Domingo. *Subject pronoun expression in Spanish: A cross-dialectal perspective*, ed. by Ana M. Carvalho, Rafael Orozco, and Naomi Lapidus Shin, 3–16. Georgetown: Georgetown University Press.

Andraszewicz, Sandra; Benjamin Scheibehenne; Jörg Rieskamp; Raoul Grasman; Josine Verhagen; and Eric-Jan Wagenmakers. 2015. An introduction to Bayesian hypothesis testing for management research. *Journal of Management* 41.521–543.

Anglim, Jeromy; Sharon Horwood; Luke D Smillie; Rosario J Marrero; and Joshua K Wood. 2020. Predicting psychological and subjective well-being from personality: A meta-analysis. *Psychological Bulletin* 146.279.

Ariel, Mira. 1990. *Accessing noun-phrase antecedents.* London: Routledge.

Ariel, Mira. 2001. Accessibility theory: An overview. *Text representation: Linguistic and psycholinguistic aspects*, ed. by Ted Sanders, Joost Schilperoord, and Wilbert Spooren, 29–87. Amsterdam: John Benjamins.

Arnold, Jennifer E. 1998. *Reference form and discourse patterns.* Stanford University dissertation.

Arnold, Jennifer E. 2001. The effect of thematic roles on pronoun use and frequency of reference continuation. *Discourse Processes* 31.137–162.

Arnold, Jennifer E. 2010. How speakers refer: The role of accessibility. *Language and Linguistics Compass* 4.187–203.

Arnold, Jennifer E, and Zenzi M Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language* 56.521–536.

Aylett, Matthew, and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech* 47.31–56.

Balduzzi, Sara; Gerta Rücker; and Guido Schwarzer. 2019. How to perform a meta-analysis with R: a practical tutorial. *BMJ Ment Health* 22.153–160. Online: `https://mentalhealth.bmj.com/content/22/4/153`.

Bell, Alan; Jason M Brenier; Michelle Gregory; Cynthia Girand; and Dan Jurafsky. 2009. Predictability effects on durations of content and function words in conversational English. *Journal of Memory and Language* 60.92–111.

Betancourt, Michael. 2017. A conceptual introduction to Hamiltonian Monte Carlo. *arXiv*, Retrieved from `https://arxiv.org/abs/1701.02434`.

Bochynska, Agata; Liam Keeble; Caitlin Halfacre; Joseph V Casillas; Irys-Amélie Champagne; Kaidi Chen; Melanie Röthlisberger; Erin M Buchanan; and Timo Roettger. 2023. Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics* 2.1–36.

Bockrath, Margaret F; Kenneth I Pargament; Serena Wong; Valencia A Harriott; Julie M Pomerleau; Steffany J Homolka; Zyad B Chaudhary; and Julie J Exline. 2022. Religious and spiritual struggles and their links to psychological adjustment: A meta-analysis of longitudinal studies. *Psychology of Religion and Spirituality* 14.283–299.

Bott, Oliver; Torgrim Solstad; and Anna Pryslopska. 2018. Implicit causality affects the choice of anaphoric form. *Poster presented at Architectures and Mechanisms for Language Processing (AMLaP)*, Berlin, Germany.

Brennan, Susan E. 1995. Centering attention in discourse. *Language and Cognitive Processes* 10.137–167.

Brennan, Susan E.; Marilyn W. Friedman; and Carl J. Pollard. 1987. A centering approach to pronouns. *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, 155–162. Stanford, California, USA:

Association for Computational Linguistics. Online: https://aclanthology.org/P87-1022.

Brocher, Andreas; Sofiana Iulia Chiriacescu; and Klaus von Heusinger. 2018. Effects of information status and uniqueness status on referent management in discourse comprehension and planning. *Discourse Processes* 55.346–370.

Brown, Alexander L; Taisuke Imai; Ferdinand M Vieider; and Colin F Camerer. 2024. Meta-analysis of empirical estimates of loss aversion. *Journal of Economic Literature* 62.485–516.

Bubic, Andreja; D Yves von Cramon; and Ricarda I Schubotz. 2010. Prediction, cognition and the brain. *Frontiers in Human Neuroscience* 4.25.

Bürki, Audrey; Shereen Elbuy; Sylvain Madec; and Shravan Vasishth. 2020. What did we learn from forty years of research on semantic interference? A Bayesian meta-analysis. *Journal of Memory and Language* 114.104–125.

Bürki, Audrey; Emiel van den Hoven; Niels Schiller; and Nikolay Dimitrov. 2023. Cross-linguistic differences in gender congruency effects: Evidence from meta-analyses. *Journal of Memory and Language* 131.104428.

Bürkner, Paul-Christian. 2021. Bayesian Item Response Modeling in R with brms and Stan. *Journal of Statistical Software* 100.1–54.

Casula, Matteo; Alessandro Andreis; Stefano Avondo; Matteo Pio Vaira; and Massimo Imazio. 2022. Colchicine for cardiovascular medicine: a systematic review and meta-analysis. *Future Cardiology* 18.647–659.

Chafe, Wallace L. 1994. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing.* Chicago, IL: University of Chicago Press.

Chamorro, Gloria. 2018. Offline interpretation of subject pronouns by native speakers of Spanish. *Glossa: a journal of general linguistics* 3.27.

Cheng, Wei, and Amit Almor. 2019. A Bayesian approach to establishing coreference in second language discourse: Evidence from implicit causality and consequentiality verbs. *Bilingualism: Language and Cognition* 22.456–475.

Chiriacescu, Sofiana, and Klaus von Heusinger. 2010. Discourse prominence and pe-marking in Romanian. *International Review of Pragmatics* 2.298–332.

Choi, Kiyong. 2013. Hankwukeuy 3 inching cisi phyohyen 'ku'ey kwuanhan soko (A note on a 3rd person referring expression ku in Korean). *Studies in Generative Grammar* 23.527–558.

Chowdhury, Mohammad Ziaul Islam; Iffat Naeem; Hude Quan; Alexander A Leung; Khokan C Sikdar; Maeve O'Beirne; and Tanvir C Turin. 2020. Summarising and synthesising regression coefficients through systematic review and meta-analysis for improving hypertension prediction using metamodelling: Protocol. *BMJ Open* 10.e036388.

Clark, Andy. 2013. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36.181–204.

Cohen, Jacob. 1988. *Statistical power analysis for the behavioral sciences*. 2nd edn. Hillsdale, NJ: Lawrence Erlbaum Associates.

Contemori, Carla, and Elisa Di Domenico. 2021. Microvariation in the division of labor between null-and overt-subject pronouns: the case of Italian and Spanish. *Applied Psycholinguistics* 42.997–1028.

Crawley, Rosalind A; Rosemary J Stevenson; and David Kleinman. 1990. The use of heuristic strategies in the interpretation of pronouns. *Journal of Psycholinguistic Research* 19.245–264.

Demberg, Vera; Ekaterina Kravtchenko; and Jia E Loy. 2023. A systematic evaluation of factors affecting referring expression choice in passage completion tasks. *Journal of Memory and Language* 130.104413.

Egger, Matthias; George Davey Smith; Martin Schneider; and Christoph Minder. 1997. Bias in meta-analysis detected by a simple, graphical test. *BMJ* 315.629–634. Online: `https://www.bmj.com/content/315/7109/629`.

Fedele, Emily, and Elsi Kaiser. 2015. Resolving null and overt pronouns in Italian. *Proceedings of the 15th texas linguistic society*, ed. by Christopher Brown, Qianping Gu, Cornelia Loos, Jason Mielens, and Grace Neveu, Austin, TX: University of Texas, 53–72.

Ferretti, Todd R; Hannah Rohde; Andrew Kehler; and Melanie Crutchley. 2009. Verb aspect, event structure, and coreferential processing. *Journal of Memory and Language* 61.191–205.

Ferstl, Evelyn C; Alan Garnham; and Christina Manouilidou. 2011. Implicit causality bias in English: A corpus of 300 verbs. *Behavior Research Methods* 43.124–135.

Filiaci, Francesca; Antonella Sorace; and Manuel Carreiras. 2014. Anaphoric biases of null and overt subjects in Italian and Spanish: a cross-linguistic comparison. *Language, Cognition and Neuroscience* 29.825–843.

Frank, Stefan L.; Leun J. Otten; Giulia Galli; and Gabriella Vigliocco. 2013. Word surprisal predicts N400 amplitude during reading. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ed. by Hinrich Schuetze, Pascale Fung, and Massimo Poesio, 878–883. Sofia, Bulgaria: Association for Computational Linguistics. Online: `https://aclanthology.org/P13-2152`.

Frederiksen, Anne Therese, and Rachel I Mayberry. 2022. Pronoun production and comprehension in American Sign Language: the interaction of space, grammar, and semantics. *Language, Cognition and Neuroscience* 37.80–102.

Fukumura, Kumiko, and Roger PG van Gompel. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language* 62.52–66.

Fukumura, Kumiko, and Roger PG van Gompel. 2011. The effect of animacy on the choice of referring expression. *Language and Cognitive Processes* 26.1472–1504.

Fukumura, Kumiko, and Roger PG van Gompel. 2015. Effects of order of mention and grammatical role on anaphor resolution. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 41.501.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. United Kingdom: Cambridge University Press.

Gelman, Andrew; Daniel Lee; and Jiqiang Guo. 2015. Stan: A probabilistic programming language for Bayesian inference and optimization. *Journal of Educational and Behavioral Statistics* 40.530–543.

Givón, Talmy. 1983. Topic continuity in discourse: An introduction. *Topic continuity in discourse: A quantitative cross-language study*, ed. by Talmy Givón, 1–42. Amsterdam: John Benjamins Publishing.

Grosz, Barbara J; Scott Weinstein; and Aravind K Joshi. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21.203–225.

Grüter, Theres; Hannah Rohde; and Amy J Schafer. 2017. Coreference and discourse coherence in L2: The roles of grammatical aspect and referential form. *Linguistic Approaches to Bilingualism* 7.199–229.

Guan, Shuang, and Jennifer E Arnold. 2021. The predictability of implicit causes: testing frequency and topicality explanations. *Discourse Processes*, 1–27.

Gundel, Jeanette K; Nancy Hedberg; and Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language*, 274–307.

Harari, Michael B; Heather R Parola; Christopher J Hartwell; and Amy Riegelman. 2020. Literature searches in systematic reviews and meta-analyses: A review, evaluation, and recommendations. *Journal of Vocational Behavior* 118.103377.

Hartshorne, Joshua K; Timothy J O'Donnell; and Joshua B Tenenbaum. 2015. The causes and consequences explicit in verbs. *Language, Cognition and Neuroscience* 30.716–734.

Heimberger, Philipp. 2020. Does economic globalisation affect income inequality? A meta-analysis. *The World Economy* 43.2960–2982.

Higgins, Julian PT; James Thomas; Jacqueline Chandler; Miranda Cumpston; Tianjing Li; Matthew J Page; and Vivian A Welch. 2019. *Cochrane handbook for systematic reviews of interventions*. Chichester, UK: John Wiley & Sons, Ltd.

Holler, Anke, and Katja Suckow. 2016. How clausal linking affects noun phrase salience in pronoun resolution. *Empirical perspectives on anaphora resolution*, ed. by Anke Holler and Katja Suckow, 61–85. Berlin: De Gruyter.

Hwang, Heeju. 2022. The influence of discourse continuity on referential form choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49.626–641.

Hwang, Heeju. 2023. Choice of nominative and topic markers in Korean discourse. *Quarterly Journal of Experimental Psychology* 76.905–921.

Hwang, Heeju, and Suet Ying Lam. 2023. The influence of action continuity on reference form in Mandarin and English. Poster presented at the 36th Annual Conference on Human Sentence Processing.

Hwang, Heeju; Suet Ying Lam; Wenjing Ni; and He Ren. 2022. The role of grammatical role and thematic role predictability in reference form production in Mandarin Chinese. *Frontiers in Psychology* 13.930572.

In'nami, Yo, and Rie Koizumi. 2009. A meta-analysis of test format effects on reading and listening test performance: Focus on multiple-choice and open-ended formats. *Language Testing* 26.219–244.

Jaeger, T Florian, and Esteban Buz. 2017. Signal reduction and linguistic encoding. *The handbook of psycholinguistics*, ed. by Eva M. Fernández and Helen Smith Cairns, 38–81. Hoboken, NJ: John Wiley & Sons Ltd.

Jäger, Lena A; Felix Engelmann; and Shravan Vasishth. 2017. Similarity-based interference in sentence comprehension: Literature review and Bayesian meta-analysis. *Journal of Memory and Language* 94.316–339.

Johnson, Elyce D, and Jennifer E Arnold. 2023. The frequency of referential patterns guides pronoun comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 49.1325.

Jurafsky, Daniel; Alan Bell; Michelle Gregory; and William D Raymond. 2001. Probabilistic relations between words: Evidence from reduction in lexical production. *Frequency and the emergence of linguistic structure*, ed. by J Bybee and P Hopper, 229–254. Amsterdam: John Benjamins.

Kaiser, Elsi. 2010. Investigating the consequences of focus on the production and comprehension of referring expressions. *International Review of Pragmatics* 2.266–297.

Kehler, Andrew; Laura Kertz; Hannah Rohde; and Jeffrey L Elman. 2008. Coherence and coreference revisited. *Journal of Semantics* 25.1–44.

Kehler, Andrew, and Hannah Rohde. 2013. A probabilistic reconciliation of coherence-driven and centering-driven theories of pronoun interpretation. *Theoretical Linguistics* 39.1–37.

Kehler, Andrew, and Hannah Rohde. 2019. Prominence and coherence in a Bayesian theory of pronoun interpretation. *Journal of Pragmatics* 154.63–78.

Kim, Haeyeon. 1990. Continuity of action and topic in discourse. *Japanese/Korean Linguistics: Volume 1*, ed. by Hajime Hoji, 79–96. Stanford, CA: Stanford University Press.

Konuk, Gökben, and Klaus von Heusinger. 2021. Discourse prominence in Turkish: The interaction of grammatical function and semantic role. *Proceedings of the 15th Workshop on Altaic Formal Linguistics (WAFL 15)*, ed. by Julia Sinitsyna and Sergei Tatevosov, *MIT Working Papers in Linguistics*, vol. 93, 109–120. Cambridge, MA: Department of Linguistics, Massachusetts Institute of Technology.

Kravtchenko, Ekaterina. 2022. *Integrating pragmatic reasoning in an efficiency-based theory of utterance choice*. Universität des Saarlandes dissertation.

Kuperberg, Gina R, and T Florian Jaeger. 2016. What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience* 31.32–59.

Lam, Suet-Ying, and Heeju Hwang. 2022. How does topicality affect the choice of referential form? evidence from Mandarin. *Cognitive Science* 46.e13190.

Langlois, Valerie J; Sandra A Zerkle; and Jennifer E Arnold. 2023. Does referential expectation guide both linguistic and social constraints on pronoun comprehension? *Journal of Memory and Language* 129.104401.

Lau, Sin Hang, and Heeju Hwang. 2016. The effects of frequency on pronoun production. *Journal of Cognitive Science* 17.547–569.

Lemoine, Nathan P. 2019. Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos* 128.912–928.

Levy, Roger, and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. *Proceedings of the 19th International Conference on Neural Information Processing Systems*, vol. 19, 849–856. Cambridge, MA, USA: MIT Press.

Liao, Xixian. 2022. Coherence-driven predictability and referential form: evidence from English corpus data. *Proceedings of Sinn und Bedeutung 26*, ed. by Daniel & Sophie Repp Gutzmann, Universität zu Köln, Germany, 544–556.

Liao, Xixian; Gemma Boleda; Hannah Rohde; and Laia Mayol. 2024. Comparing models of pronoun production and interpretation via observational and experimental evidence. *Glossa: a journal of general linguistics* 9.1–46.

Liao, Xixian; Laia Mayol; and Hannah Rohde. 2023. Bayesian pronoun interpretation in English corpus passage completions. Poster presented at the 36th Annual Conference on Human Sentence Processing.

Lindemann, Sofiana-Iulia; Stanca Mada; Laura Sasu; and Madalina Matei. 2020. Thematic role and grammatical function affect pronoun production. *Proceedings of 11th International Conference of Experimental Linguistics (ExLing 2020)*, ed. by Antonis Botinis, Athens, Greece: International Society of Experimental Linguistics, 105–108.

Mahowald, Kyle; Ariel James; Richard Futrell; and Edward Gibson. 2016. A meta-analysis of syntactic priming in language production. *Journal of Memory and Language* 91.5–27.

Mayol, Laia. 2018. Asymmetries between interpretation and production in Catalan pronouns. *Dialogue & Discourse* 9.1–34.

McCoy, Kathleen F, and Michael Strube. 1999. Generating anaphoric expressions: Pronoun or definite description? *Proceedings of the Workshop on Relation of Discourse/Dialogue Structure and Reference*, ed. by Dan Cristea, Nancy Ide, and Daniel Marcu, College Park, MD, USA: University of Maryland, 65–71.

McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and STAN (2nd ed.)*. New York, USA: Chapman and Hall/CRC.

Medina Fetterman, Ana M; Natasha N Vazquez; and Jennifer E Arnold. 2022. The effects of semantic role predictability on the production of overt pronouns in Spanish. *Journal of Psycholinguistic Research* 51.169–194.

Modi, Ashutosh; Ivan Titov; Vera Demberg; Asad Sayeed; and Manfred Pinkal. 2017. Modeling semantic expectation: Using script knowledge for referent prediction. *Transactions of the Association for Computational Linguistics* 5.31–44.

Moens, Marc, and Mark Steedman. 1988. Temporal ontology and temporal reference. *Computational Linguistics* 14.15–28. Online: `https://aclanthology.org/J88-2003`.

Morris, Paul; Chong Wang; and Annette O'Connor. 2024. Network meta-analysis for an ordinal outcome when outcome categorization varies across trials. *Systematic Reviews* 13.128. Online: `https://doi.org/10.1186/s13643-024-02537-w`.

Nicenboim, Bruno; Timo B Roettger; and Shravan Vasishth. 2018. Using meta-analysis for evidence synthesis: The case of incomplete neutralization in German. *Journal of Phonetics* 70.39–55.

Nicenboim, Bruno; Daniel Schad; and Shravan Vaishth. 2024. An introduction to Bayesian data analysis for cognitive science. Version October 2024. Available at `https://bruno.nicenboim.me/bayescogsci/`.

Özge, Umut; Duygu Özge; and Klaus von Heusinger. 2016. Strong indefinites in Turkish, referential persistence, and salience structure. *Empirical perspectives on anaphora resolution*, ed. by Anke Holler and Katja Suckow, 169–191. Berlin: De Gruyter.

Patterson, Clare; Petra B Schumacher; Bruno Nicenboim; Johannes Hagen; and Andrew Kehler. 2022. A Bayesian approach to German personal and demonstrative pronouns. *Frontiers in Psychology* 12.6296.

Perret, Cyril, and Patrick Bonin. 2019. Which variables should be controlled for to investigate picture naming in adults? A Bayesian meta-analysis. *Behavior Research Methods* 51.2533–2545. Online: `https://doi.org/10.3758/s13428-018-1100-1`.

Piantadosi, Steven T; Harry Tily; and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition* 122.280–291.

Portele, Yvonne, and Markus Bader. 2020. Coherence and the interpretation of personal and demonstrative pronouns in German. *Information structuring in discourse*, ed. by Anke Holler, Katja Suckow, and Israel de la Fuente, 24–55. Leiden: Brill.

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

Riley, Richard D; Karel G M Moons; Kym I E Snell; Joie Ensor; Lotty Hooft; Douglas G Altman; Jill Hayden; Gary S Collins; and Thomas P A Debray. 2019. A guide to systematic review and meta-analysis of prognostic factor studies. *BMJ* 364.k4597.

Rohde, Hannah. 2008. *Coherence-driven effects in sentence and discourse processing.* University of California, San Diego dissertation. Online: `https://www.proquest.com/dissertations-theses/coherence-driven-effects-sentence-discourse/docview/304659507/se-2`.

Rohde, Hannah, and Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience* 29.912–927.

Rosa, Elise C. 2015. *Semantic role predictability affects referential form.* The University of North Carolina at Chapel Hill dissertation.

Rosa, Elise C, and Jennifer E Arnold. 2017. Predictability affects production: Thematic roles can affect reference form selection. *Journal of Memory and Language* 94.43–60.

Smith, Nathaniel J, and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition* 128.302–319.

Solstad, Torgrim, and Oliver Bott. 2022. On the nature of implicit causality and consequentiality: the case of psychological verbs. *Language, Cognition and Neuroscience* 27.1–30.

Sonderegger, Morgan. 2023. *Regression modeling for linguistic data.* Cambridge, MA: MIT Press.

Staub, Adrian. 2015. The effect of lexical predictability on eye movements in reading: Critical review and theoretical interpretation. *Language and Linguistics Compass* 9.311–327.

Stevenson, Rosemary J; Rosalind A Crawley; and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and Cognitive Processes* 9.519–548.

Thompson, Christopher G, and Brandie Semma. 2020. An alternative approach to frequentist meta-analysis: A demonstration of Bayesian meta-analysis in adolescent development research. *Journal of Adolescence* 82.86–102.

Tily, Harry, and Steven Piantadosi. 2009. Refer efficiently: Use less informative expressions for more predictable meanings. *Proceedings of the workshop on the production of referring expressions: Bridging the gap between computational and empirical approaches to reference (PRE-CogSci 2009)*, ed. by Kees van Deemter, Albert Gatt, Roger van Gompel, and Emiel Krahmer, Amsterdam: Cognitive Science Society.

Torregrossa, Jacopo; Maria Andreou; and Christiane M Bongartz. 2020. Variation in the use and interpretation of null subjects: A view from Greek and Italian. *Glossa: a journal of general linguistics* 5.95.

Vehtari, Aki; Andrew Gelman; and Jonah Gabry. 2017. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27.1413–1432. Online: https://doi.org/10.1007/s11222-016-9696-4.

Vehtari, Aki; Andrew Gelman; Daniel Simpson; Bob Carpenter; and Paul-Christian Bürkner. 2021. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis* 16.667–718.

Verhagen, Véronique; Maria Mos; Ad Backus; and Joost Schilperoord. 2018. Predictive language processing revealing usage-based variation. *Language and Cognition* 10.329–373.

Vogels, Jorrig. 2019. Both thematic role and next-mention biases affect pronoun use in Dutch. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 3029–3035.

Wang, Ying; Liang Chen; Wenhua Xiang; Shuai Ouyang; Taidong Zhang; Xiulan Zhang; Yelin Zeng; Yanting Hu; Gongwen Luo; and Yakov Kuzyakov. 2021. Forest conversion to plantations: A meta-analysis of consequences for soil and microbial properties and functions. *Global Change Biology* 27.5643–5656.

Weatherford, Kathryn C, and Jennifer E Arnold. 2021. Semantic predictability of implicit causality can affect referential form choice. *Cognition* 214.104759.

Wu, Nicholas C, and Frank Seebacher. 2020. Effect of the plastic pollutant bisphenol A on the biology of aquatic organisms: A meta-analysis. *Global Change Biology* 26.3821–3833.

Xing, Qi; Ling Fu; Zhichao Yu; and Xueping Zhou. 2020. Efficacy and safety of integrated traditional Chinese medicine and Western medicine on the treatment of rheumatoid arthritis: a meta-analysis. *Evidence-Based Complementary and Alternative Medicine* 2020.4348709.

Ye, Yining, and Jennifer Arnold. 2023. Implicit causality affects pronoun use for speakers (but not writers). Poster presented at the 36th Annual Conference on Human Sentence Processing.

Ye, Yining, and Jennifer Arnold. 2025. Implicit causality can affect pronoun use in interactive fragment completion tasks. *Journal of Experimental Psychology: Learning, Memory and Cognition*, to appear.

Yousefi, Mohammad Hossein, and Reza Biria. 2018. The effectiveness of L2 vocabulary instruction: a meta-analysis. *Asian-Pacific Journal of Second and Foreign Language Education* 3.21.

Zerkle, Sandra A, and Jennifer E Arnold. 2019. Does pre-planning explain why predictability affects reference production? *Dialogue & Discourse* 10.34–55.

Zhan, Meilin; Roger Levy; and Andrew Kehler. 2020. Pronoun interpretation in Mandarin Chinese follows principles of Bayesian inference. *PLOS ONE* 15.e0237012.

NOTES

1 Note that the aspect of predictability that is at stake here is referent predictability, construed as the addressee's estimate of the likelihood that the speaker will mention a given referent in the upcoming discourse.

2 Abbreviations in Example 8: NOM = nominative, ACC = accusative.

3 URL: `https://scholar.google.com/`

4 Arnold 1998 and Arnold 2001 report on the same experiment. We have included both references because studies in this field may cite either or both of them.

5 Second language learners have also been investigated on this matter. For instance, research has been conducted on Japanese- and Korean-speaking learners of English (Grüter et al. 2017), as well as Chinese-speaking English learners (Cheng & Almor 2019).

6 By the time of manuscript submission, the study had been published and was searchable.

7 We conducted a sensitivity analysis assessing the effect of using our chosen priors compared to the *brms* default priors for fixed effects. We found no major differences in the results. Details of this analysis are reported in Appendix C.

8 In *brms* syntax, this corresponds to $formula = estimate \mid se(error) \sim 1 + (1 \mid article/sample)$. The fixed intercept item 1 represents the estimated effect over studies; while the term *(1 | article / sample)* allows for heterogeneity between studies and between samples nested within studies.

9 In the analyses, languages were classified according to their respective language families, primarily due to the limited availability of data for individual languages. To ensure the validity of our subsequent results, we performed a sensitivity analysis, using Language as a covariate instead of Language family. As expected, using Language results in greater uncertainty concerning the summary effect estimate, with the general pattern of the estimates unchanged. We report this analysis in Appendix H.5.

10 The study by Ye and Arnold (2023) corresponds to the first two experiments reported in the article by Ye and Arnold (2025). A third experiment in Ye & Arnold 2025 tested a written task that was interactive with a live addressee, suggesting that interactivity, rather than modality, may account for the observed differences in the earlier experiments. Our focus on modality as a factor was based on the findings in Ye & Arnold 2023, as this was the available evidence at the time we conducted our analyses.

11 As previously mentioned in Section 3.3, while some English studies included zero pronouns as part of the dependent measure, these are not null pronouns per se. Our current analysis does not cover zero pronouns, and we consider investigating this in future research.

12 The estimate of 0.28 [-0.00, 0.56] in log odds from the model was converted to an odds ratio 1.32 [1, 1.74, 95% CI]. For the sake of interpretability, all the effect sizes presented in this study have been transformed in this way.

13 We used *brms'* function *hypothesis()* for this hypothesis test. Bayes Factors quantify the strength of evidence for one hypothesis over another. Following Andraszewicz et al. 2015, BFs between 30 and 100 indicate very strong evidence for the alternative hypothesis, while values between 10 and 30 indicate strong evidence, values between 3 and 10 indicate moderate evidence, and values between 1 and 3 indicate anecdotal evidence. BFs below 1 reflect varying levels of evidence for the null hypothesis, with smaller values indicating stronger evidence for the null. One could also use the probability $p(\beta>0)$ (i.e., p(odds ratio$>$1)) instead. In our case, the effect of

predictability has a 97.25% probability of being positive. However, $p(\beta>0)$ and BFs reflect the same information in different forms (probability versus odds). In what follows, we will report only the BF, as it is commonly used in Bayesian hypothesis testing and provides an intuitive interpretation of evidence strength.

14 The number of datapoints by each predictor and each level is reported in Appendix G.

15 The fuller model, which included an interaction between manipulation type and language family as a predictor, failed to converge. It had saturated trajectories and indicated a lack of chain mixing even after increasing the maximum tree-depth and the number of iterations the chains ran for. This is likely due to the sparsity of data for certain combinations for this interaction. Some combinations, such as Turkic under Relative Clause and Relation, and Korean under TPV for object referents, have no data at all. See Table 17 in Appendix G for a detailed summary of the data structure.

16 We excluded Korean overt pronouns from our analysis due to their rarity and similarity to noun phrases (Kim 1990, Choi 2013, Hwang 2023). Retaining Korean data for null pronouns poses no harm, as the data is partially pooled across languages, contributing evidence solely for the null-subject instances; however, care is advised when interpreting language-level predictors for Korean.

17 For readers interested in the analysis that does not control for potential variations from the four additional predictors or their interactions, the estimated overall effect of predictability is 1.31 [0.73, 2.51, 95%CI]. The Bayesian hypothesis test provides support for a positive effect of predictability, with a BF of 5.

xixianliao@gmail.com thomas.brochhagen@upf.edu
gemma.boleda@upf.edu laia.mayol@upf.edu

7. List of figures.

1 Flow diagram showing study selection for the meta-analysis. 63
2 Diagram illustrating the two stages of the analysis. 64
3 Forest plot illustrating the estimated distribution of the difference in the use of the most reduced referential form between referents of higher and lower predictability. Effect estimates are represented by black dots, while the horizontal lines depict 95% credible intervals (CIs). An odds ratio greater than 1 indicates evidence supporting a positive effect of predictability on pronoun production. The top-most row corresponds to the overall estimate obtained from collating the evidence from the individual studies below it. The vertical bold black line represents an odds ratio of 1, signifying no difference, while the thick dashed line denotes the mean of the overall estimate. The two thinner dashed lines represent the 95% CI of the overall effect. The CI for Hwang (2022b) is much wider and has been truncated to reduce excessive white space in the plot. 65
4 Forest plot visualizing the overall effect and the estimated odds ratios of the individual experiments after accounting for potential variations from grammatical function the antecedent, manipulation type and language family. The CI for Hwang (2022b) is wider and has been truncated to reduce excessive white space in the plot. 66
5 Estimated effect size of predictability (in log odds) showing interaction between language family and grammatical function of the antecedent. Posterior medians are shown with 95% credible intervals, averaging over manipulation type. 67
6 Forest plot visualizing the overall effect of predictability on null-subject languages and the estimated odds ratios of the individual experiments. The CI for Hwang (2022b) is wider and has been truncated to reduce excessive white space in the plot. 68
7 Prior predictive check. The plot shows the observed data (y, dark line) and 100 simulated datasets (y_{rep}, light lines) generated from the prior-only model. The x-axis represents the predicted probability of pronoun production (0 to 1), and the y-axis represents the density. 69
8 Funnel plot. 70

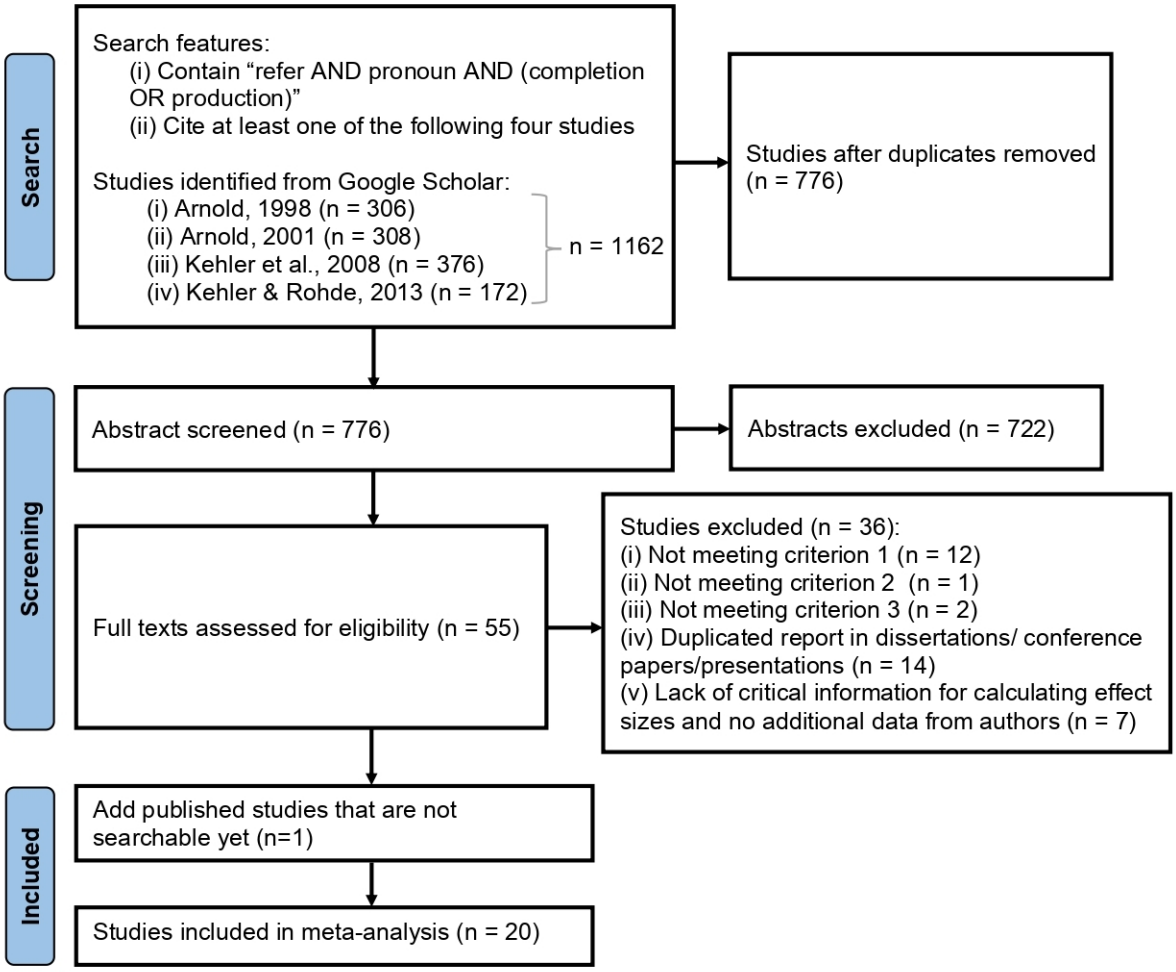

FIGURE 1. Flow diagram showing study selection for the meta-analysis.

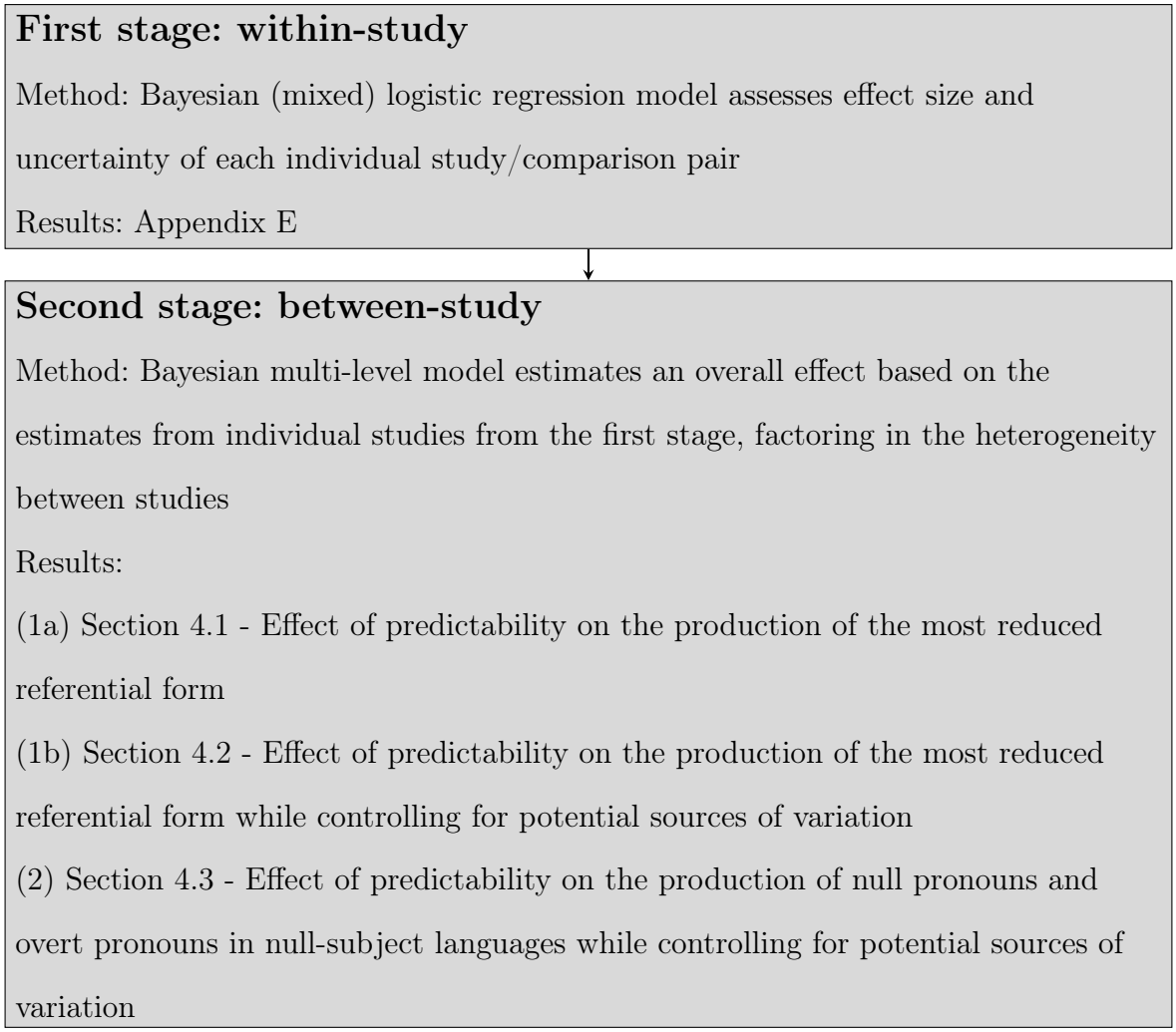

FIGURE 2. Diagram illustrating the two stages of the analysis.

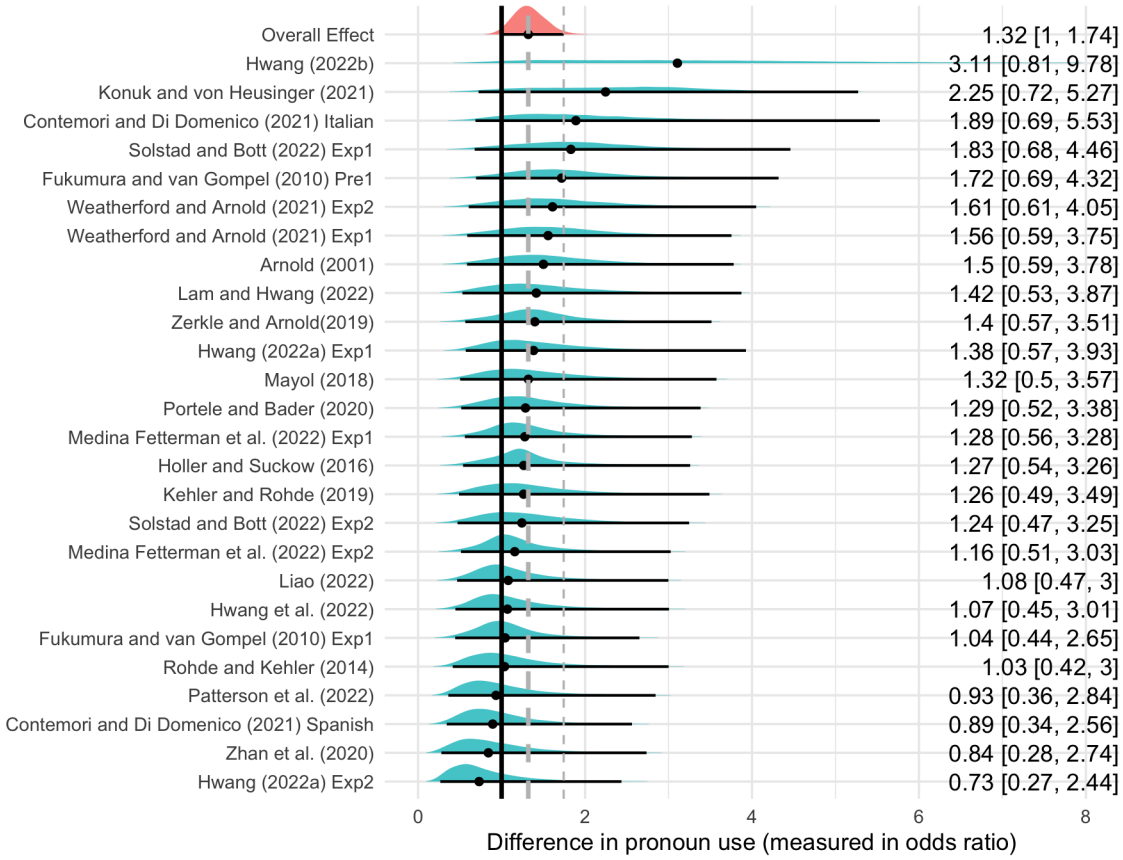

FIGURE 3. Forest plot illustrating the estimated distribution of the difference in the use of the most reduced referential form between referents of higher and lower predictability. Effect estimates are represented by black dots, while the horizontal lines depict 95% credible intervals (CIs). An odds ratio greater than 1 indicates evidence supporting a positive effect of predictability on pronoun production. The top-most row corresponds to the overall estimate obtained from collating the evidence from the individual studies below it. The vertical bold black line represents an odds ratio of 1, signifying no difference, while the thick dashed line denotes the mean of the overall estimate. The two thinner dashed lines represent the 95% CI of the overall effect. The CI for Hwang (2022b) is much wider and has been truncated to reduce excessive white space in the plot.

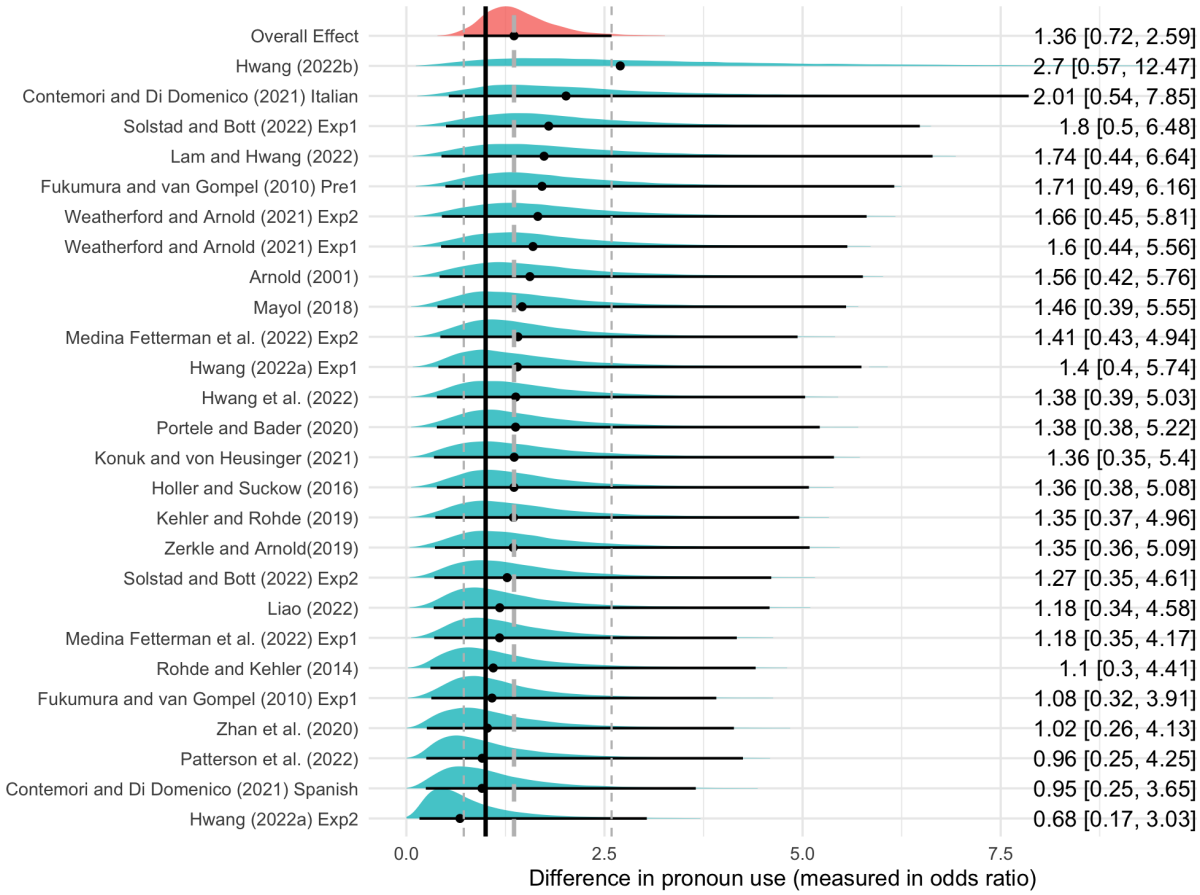

FIGURE 4. Forest plot visualizing the overall effect and the estimated odds ratios of the individual experiments after accounting for potential variations from grammatical function the antecedent, manipulation type and language family. The CI for Hwang (2022b) is wider and has been truncated to reduce excessive white space in the plot.

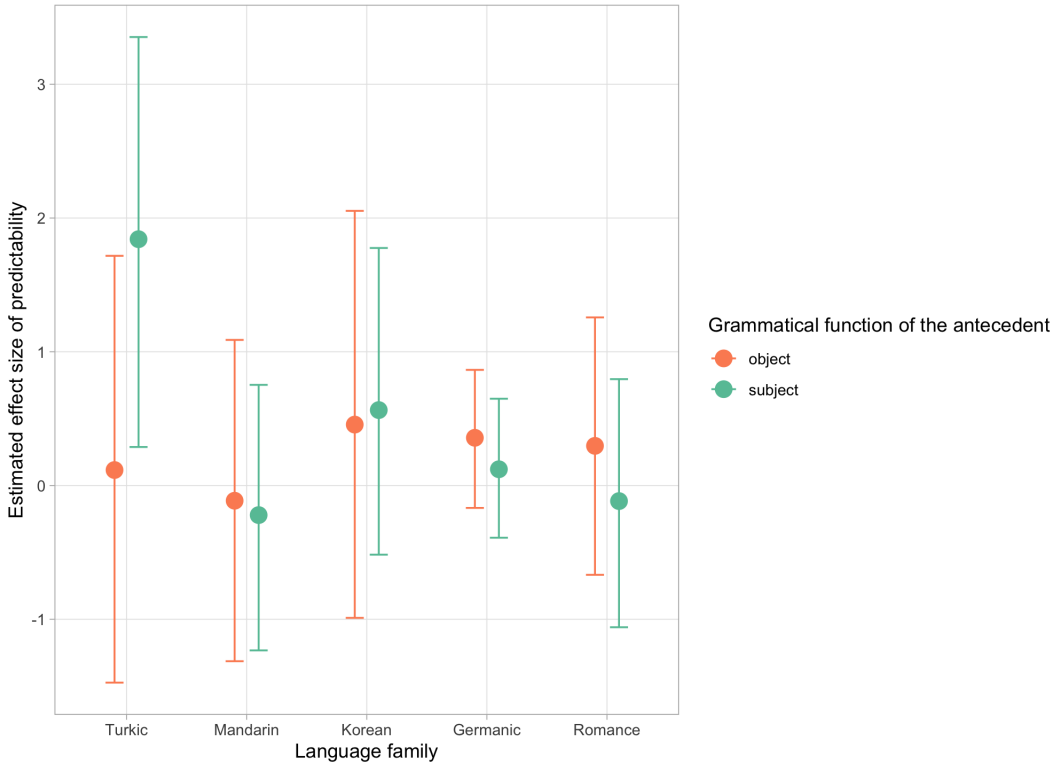

FIGURE 5. Estimated effect size of predictability (in log odds) showing interaction between language family and grammatical function of the antecedent. Posterior medians are shown with 95% credible intervals, averaging over manipulation type.

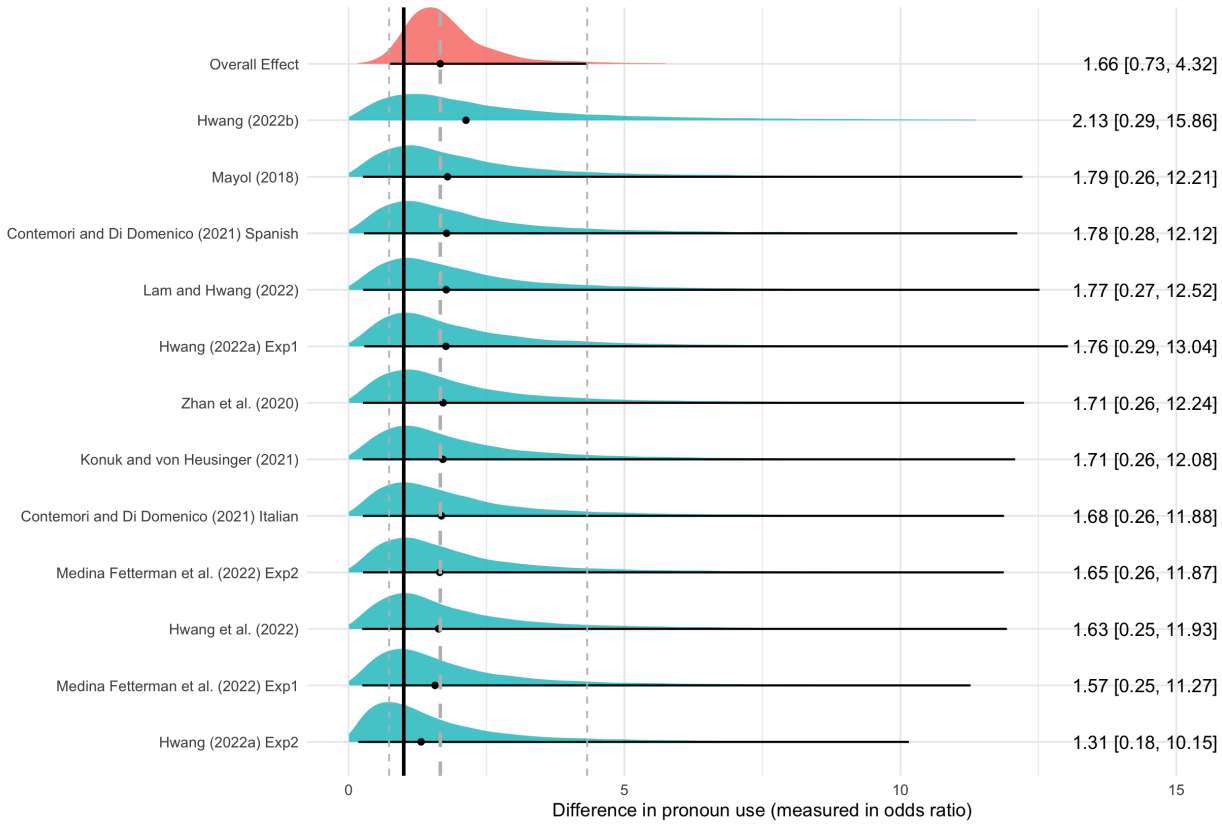

FIGURE 6. Forest plot visualizing the overall effect of predictability on null-subject languages and the estimated odds ratios of the individual experiments. The CI for Hwang (2022b) is wider and has been truncated to reduce excessive white space in the plot.

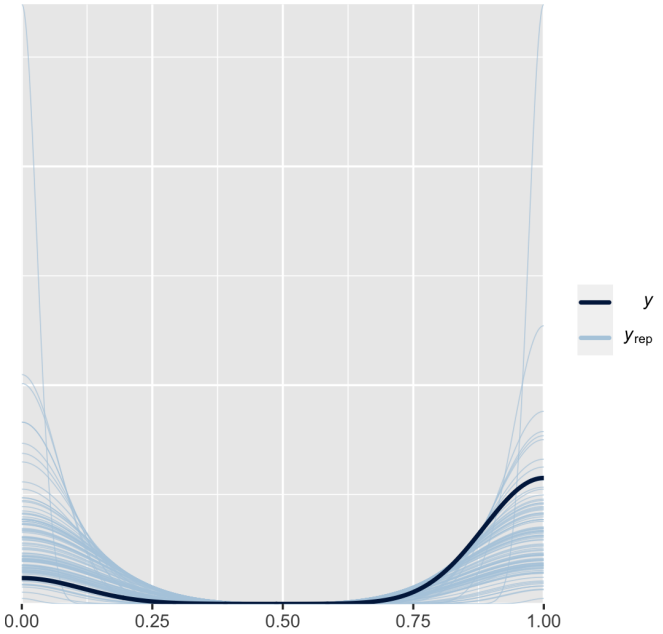

FIGURE 7. Prior predictive check. The plot shows the observed data (y, dark line) and 100 simulated datasets (y_{rep}, light lines) generated from the prior-only model. The x-axis represents the predicted probability of pronoun production (0 to 1), and the y-axis represents the density.

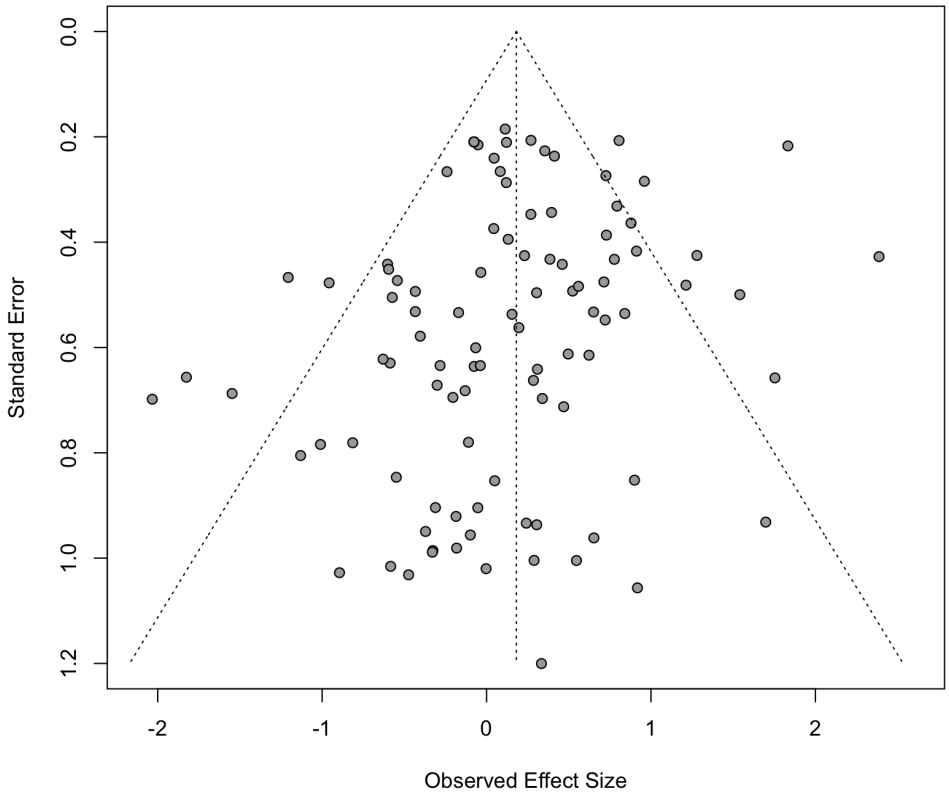

FIGURE 8. Funnel plot.

8. List of tables.

1 Overview of the samples included in the meta-analysis, with the manipulation type used and the conclusion drawn in each study. TPV: transfer-of-possession verbs. ICV: implicit causality verbs. 73
2 Example of contingency table with made-up data. Cells show number of referring expressions produced for the more predictable referent (A and C) and the less predictable one (B and D). 74
3 Effect estimate (in log odds) of predictability on the use of the most reduced reference form. This table summarizes the model incorporating three predictors and their interactions: manipulation type, language family, and grammatical function of the antecedent (abbreviated as *gram_func_ant*). 75
4 Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with four covariates added (manipulation type, language family, grammatical function of the antecedent, and pronoun type), including interaction terms. 76
5 Samples included in the meta-analysis, with publication type and number of participants included in the analysis. 77
6 Overview of conclusions drawn in previous work. TPV: transfer-of-possession verbs. ICV: implicit causality verbs. ASL: American Sign Language. 78
7 Dataset availability 79
8 Comparison of model fits using two different priors for analyzing pronoun production referring to predictable versus less predictable subject referents in Rohde & Kehler 2014. Each row presents results for a parameter with both the weakly informative prior and the *brms* default prior. 80
9 Model fit using the *brms* default prior comparing the null pronoun production for more predictable object referents (subject-biased ICV + so) vs. null pronoun produced referring to less predictable object referents (subject-biased ICV + because) in Hwang et al. 2022. 81
10 Summary of the posterior distribution for each comparison pair of subject referents. The estimate represents the effect size (in log odds) for each comparison pair, estimated based on the data gathered from each experiment. 83
11 Summary of the posterior distribution for each comparison pair of object referents. The estimate represents the effect size (in log odds) for each comparison pair, estimated based on the data gathered from each experiment. 84
12 Ranking of models in terms of expected log-predictive densities (ELPD, higher is better). $ELPD_{\triangle}$ is the estimated difference in ELPDs from the highest ranked model, and $SE_{\triangle}$ is the standard error of that difference. . . . 85
13 Ranking of models in terms of expected log-predictive densities (ELPD, higher is better). $ELPD_{\triangle}$ is the estimated difference in ELPDs from the highest ranked model, and $SE_{\triangle}$ is the standard error of that difference. . . . 86
14 Intercept estimates and Bayes Factor values for all models evaluated in the sensitivity analysis. 87
15 Number of datapoints by predictor and level for the analysis in Section 4.2. . 88
16 Number of datapoints by predictor and level for the analysis in Section 4.3. . 89

17 Number of datapoints by language family, grammatical function, and manipulation type in Section 4.2. 90
18 Effect estimate (in log odds) of predictability on the use of pronouns in Germanic languages: Summary of the model with two predictors (manipulation type and grammatical function of the antecedent), including interaction terms. 91
19 Effect estimate (in log odds) of predictability on the use of the most reduced referential form: Summary of the model with three covariates (manipulation type, language family, and grammatical function of the antecedent) and interactions between them, with the Spanish data excluded. 92
20 Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with four covariates and their interactions added (manipulation type, language family, grammatical function of the antecedent, and pronoun type), with Spanish data excluded. 93
21 Effect estimate (in log odds) of predictability on the use of the most reduced reference form: summary of the model with three covariates and their interactions added (manipulation type, language family, and grammatical function of the antecedent), with Contemori & Di Domenico 2021 and Solstad & Bott 2022 excluded. 94
22 Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with four covariates and their interactions added (manipulation type, language family, grammatical function of the antecedent, and pronoun type), with Contemori & Di Domenico 2021 and Solstad & Bott 2022 excluded. 95
23 Effect estimate (in log odds) of predictability on the use of the most reduced reference form: Summary of the model with manipulation type, language, grammatical function of the antecedent and their interactions added as covariates. 96
24 Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with manipulation type, language, grammatical function of the antecedent, pronoun type and their interactions added as covariates. 97
25 Effect estimate (in log odds) of predictability on the use of the most reduced reference form in English and Spanish: Summary of the model with task modality added as the only covariate. 98
26 Effect estimate (in log odds) of predictability on the use of the most reduced reference form in non-corpus studies: Summary of the model with character gender added as the only covariate. 99
27 Effect estimate (in log odds) of predictability on the use of the most reduced reference form: Summary of the model with completion paradigm added as the only covariate. 100

	experiment	language	manipulation	conclusion
1	Arnold (2001)	English	TPV	✓
2	Fukumura & van Gompel (2010) Pre1	English	ICV	✗
3	Fukumura & van Gompel (2010) Exp1	English	ICV	✗
4	Rohde & Kehler (2014)	English	ICV	✗
5	Holler & Suckow (2016)	German	ICV, relation	✗
6	Mayol (2018)	Catalan	ICV	✗
7	Kehler & Rohde (2019)	English	relative clause	✗
8	Zerkle & Arnold (2019)	English	TPV	✓[a]
9	Portele & Bader (2020)	German	relation	✗[b]
10	Zhan et al. (2020)	Mandarin	ICV	✗
11	Contemori & Di Domenico (2021) Exp1[c]	Spanish[d]	ICV	-
12	Contemori & Di Domenico (2021) Exp2[c]	Italian	ICV	-
13	Weatherford & Arnold (2021) Exp1[e]	English	ICV	✓
14	Weatherford & Arnold (2021) Exp2[e]	English	ICV	✓
15	Konuk & von Heusinger (2021)	Turkish	ICV	✓[f]
16	Hwang et al. (2022)	Mandarin	ICV, TPV, relation	✗
17	Hwang (2022)	Korean	connective	✓
18	Hwang (2023) Exp1	Korean	ICV/TPV + relation	✗
19	Hwang (2023) Exp2	Korean	ICV, TPV	✗
20	Lam & Hwang (2022)	Mandarin	ICV	✗[g]
21	Liao (2022)[h]	English	relation	✗
22	Medina Fetterman et al. (2022) Exp1	Spanish	TPV	✓[i]
23	Medina Fetterman et al. (2022) Exp2	Spanish	TPV	✓
24	Patterson et al. (2022)	German	ICV	✗
25	Solstad & Bott (2022) Exp1[c]	German	ICV	-
26	Solstad & Bott (2022) Exp2[c]	German	ICV	-

[a] Only the subject continuation trials were analyzed as speakers rarely used reduced expressions for non-subject continuation trials (10%).

[b] The pronominalization rate was lower for the more predictable Experiencer and the less predictable Stimulus.

[c] Though not targeting our research question, we analyzed these experiments because their use of the typical experimental paradigm enabled us to calculate the effect size of predictability. We detail sensitivity analyses excluding these studies in Appendix H.4.

[d] This study examines Mexican Spanish, focusing on undergraduate students at Universidad Autónoma de Ciudad Juárez. The authors noted that the region's proximity to the U.S. may render the local Spanish a contact variety.

[e] Predictability effect was detected within objects, not subjects.

[f] Predictability effect was observed within subjects, not objects.

[g] This study reported a negative effect of predictability: participants used more null pronouns for less predictable referents.

[h] Liao 2022 is the only corpus-based study. It meets our inclusion criteria and enabling effect size calculation. However, it didn't control for some factors that potentially affecting results. We performed a sensitivity analysis excluding this study to mitigate potential biases (see Appendix H.2).

[i] This experiment found that the effect of predictability only emerged for overt pronouns when used to refer to nonsubject characters.

TABLE 1. Overview of the samples included in the meta-analysis, with the manipulation type used and the conclusion drawn in each study. TPV: transfer-of-possession verbs. ICV: implicit causality verbs.

	more predictable referent	less predictable referent
pronoun	40 (A)	30 (B)
non-pronoun	10 (C)	20 (D)

TABLE 2. Example of contingency table with made-up data. Cells show number of referring expressions produced for the more predictable referent (A and C) and the less predictable one (B and D).

	Estimate	Est. Error	95% CI	BF
Intercept	0.31	0.32	[-0.32, 0.95]	5.56
manipulationType ICV	0.03	0.25	[-0.46, 0.51]	1.22
manipulationType relativeClause	-0.15	0.64	[-1.39, 1.12]	0.67
manipulationType relation	-0.01	0.25	[-0.52, 0.47]	0.89
languageFamily Turkish	0.64	0.64	[-0.67, 1.88]	5.99
languageFamily Mandarin	-0.50	0.44	[-1.40, 0.39]	0.14
languageFamily Korean	0.20	0.53	[-0.80, 1.32]	1.73
languageFamily Germanic	-0.10	0.30	[-0.69, 0.48]	0.61
gram_func_ant object	-0.12	0.14	[-0.38, 0.15]	0.25
languageFamily Turkish:gram_func_ant object	-0.75	0.17	[-1.10, -0.42]	0.00
languageFamily Mandarin:gram_func_ant object	0.16	0.21	[-0.26, 0.59]	3.74
languageFamily Korean:gram_func_ant object	0.06	0.27	[-0.46, 0.57]	1.42
languageFamily Germanic:gram_func_ant object	0.22	0.10	[0.02, 0.42]	61.50
gram_func_ant object:manipulationType ICV	0.01	0.12	[-0.23, 0.24]	1.20
gram_func_ant object:manipulationType relativeClause	0.55	0.31	[-0.05, 1.16]	27.99
gram_func_ant object:manipulationType relation	-0.11	0.13	[-0.37, 0.14]	0.24

TABLE 3. Effect estimate (in log odds) of predictability on the use of the most reduced reference form. This table summarizes the model incorporating three predictors and their interactions: manipulation type, language family, and grammatical function of the antecedent (abbreviated as *gram_func_ant*).

	Estimate	Est. Error	95% CI	BF
Intercept	0.51	0.45	[-0.31, 1.46]	9.15
gram_func_ant object	-0.17	0.11	[-0.38, 0.03]	0.05
pronounType overt	-0.11	0.06	[-0.23, 0.02]	0.04
manipulation ICV	-0.41	0.25	[-0.90, 0.07]	0.05
manipulation relation	0.10	0.22	[-0.35, 0.51]	2.21
languageFamily Turkic	0.48	0.95	[-1.37, 2.50]	2.75
languageFamily Mandarin	-0.19	0.66	[-1.61, 1.17]	0.59
languageFamily Korean	0.06	0.81	[-1.52, 1.72]	1.13
gram_func_ant object:pronounType overt	0.15	0.06	[0.03, 0.27]	107.11
gram_func_ant object:manipulation ICV	0.26	0.12	[0.03, 0.48]	73.07
gram_func_ant object:manipulation relation	-0.24	0.17	[-0.58, 0.10]	0.09
gram_func_ant object:languageFamily Turkic	-0.64	0.17	[-0.98, -0.30]	0.00
gram_func_ant object:languageFamily Mandarin	0.18	0.16	[-0.14, 0.50]	6.35
gram_func_ant object:languageFamily Korean	0.35	0.26	[-0.15, 0.86]	10.80

TABLE 4. Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with four covariates added (manipulation type, language family, grammatical function of the antecedent, and pronoun type), including interaction terms.

	experiment	type	sample size
1	Arnold (2001)	journal	16
2	Fukumura & van Gompel (2010) Pre1	journal	24
3	Fukumura & van Gompel (2010) Exp2	journal	24
4	Rohde & Kehler (2014)	journal	28
5	Holler & Suckow (2016)	book	96
6	Mayol (2018)	journal	78
7	Kehler & Rohde (2019)	journal	40
8	Zerkle & Arnold (2019)	journal	34
9	Portele & Bader (2020)	book	32
10	Zhan et al. (2020)	journal	50
11	Contemori & Di Domenico (2021) Exp1	journal	24
12	Contemori & Di Domenico (2021) Exp2	journal	24
13	Weatherford & Arnold (2021) Exp1	journal	56
14	Weatherford & Arnold (2021) Exp2	journal	46
15	Konuk & von Heusinger (2021)	proceedings	90
16	Hwang et al. (2022)	journal	62
17	Hwang (2022)	journal	34
18	Hwang (2023) Exp1	journal	57
19	Hwang (2023) Exp2	journal	65
20	Lam & Hwang (2022)	journal	40
21	Liao (2022)	proceedings	corpus
22	Medina Fetterman et al. (2022) Exp1	journal	43[a]
23	Medina Fetterman et al. (2022) Exp2	journal	26[b]
24	Patterson et al. (2022)	journal	40
25	Solstad & Bott (2022) Exp1	journal	52
26	Solstad & Bott (2022) Exp2	journal	64

[a] Participants were from Argentina, Chile, Colombia, Costa Rica, Dominican Republic, Ecuador, El Salvador, Guatemala, Mexico, Panama, Peru, Puerto Rico, Spain, Uruguay, and Venezuela.

[b] The study involved participants from seven countries/territories, residing in the United States for 0-21 years, with one individual raised between Colombia and Argentina. 19 participants were multilingual, speaking at least one additional language.

TABLE 5. Samples included in the meta-analysis, with publication type and number of participants included in the analysis.

	study	language	manipulation	conclusion
1	Arnold (2001)	English	TPV	✓
2	Ferretti et al. (2009)[a]	English	verb aspect	✗
3	Fukumura & van Gompel (2010)	English	ICV	✗
4	Rohde & Kehler (2014)	English	ICV	✗
5	Rosa (2015)	English	TPV	inconclusive [b]
6	Holler & Suckow (2016)	German	ICV, relation	✗
7	Rosa & Arnold (2017)	English	TPV	✓
8	Mayol (2018)	Catalan	ICV	✗
9	Kehler & Rohde (2019)	English	relative clause	✗
10	Zerkle & Arnold (2019)	English	TPV	✓[c]
11	Lindemann et al. (2020)	Romanian	TPV	✓
12	Portele & Bader (2020)	German	relation	✗[d]
13	Zhan et al. (2020)	Mandarin	ICV	✗
14	Konuk & von Heusinger (2021)	Turkish	ICV	✓[e]
15	Weatherford & Arnold (2021)	English	ICV	✓[f]
16	Frederiksen & Mayberry (2022)	ASL	ICV	✗
17	Hwang (2023)	Korean	ICV, TPV, relation	✗
18	Hwang (2022)	Korean	connective	✓
19	Hwang et al. (2022)	Mandarin	ICV, TPV, relation	✗
20	Kravtchenko (2022)	English	ICV, TPV	✗
21	Lam & Hwang (2022)	Mandarin	ICV	✗[g]
22	Liao (2022)	English	relation	✗
23	Medina Fetterman et al. (2022)	Spanish	TPV	✓[h]
24	Patterson et al. (2022)	German	ICV	✗
25	Hwang & Lam (2023)	Mand, Eng	relation	✗
26	Liao et al. (2023)	English	relation	✗
27	Ye & Arnold (2023)	English	ICV	✗, ✓[i]

[a] See also Rohde 2008, Experiment VII.

[b] Experiment 3 in this study did not observe more pronouns produced for more predictable referents but speculated that this may have been due to an issue of power.

[c] Only the subject continuation trials were analyzed as speakers rarely used reduced expressions for non-subject continuation trials (10%).

[d] The pronominalization rate was lower for the more predictable Experiencer and the less predictable Stimulus.

[e] Predictability effect was observed within subjects, not objects.

[f] Predictability effect was detected within objects, not subjects.

[g] This study reported a negative effect of predictability: participants used more null pronouns for less predictable referents.

[h] This study found that the effect of predictability only emerged for overt pronouns when used to refer to nonsubject characters.

[i] This study reported a significant effect of predictability in a spoken experiment, whereas no such effect was observed in a written experiment.

TABLE 6. Overview of conclusions drawn in previous work. TPV: transfer-of-possession verbs. ICV: implicit causality verbs. ASL: American Sign Language.

No.	Experiment	Available complete dataset
1	Arnold (2001)	Yes
2	Contemori & Di Domenico (2021) Italian	Yes
3	Contemori & Di Domenico (2021) Spanish	Yes
4	Fukumura & van Gompel (2010) Pre.1	No
5	Fukumura & van Gompel (2010) Exp.1	No
6	Holler & Suckow (2016)	Yes
7	Hwang et al. (2022)	Yes
8	Hwang (2023) Exp.1	Yes
9	Hwang (2023) Exp.2	Yes
10	Hwang (2022)	Yes
11	Konuk & von Heusinger (2021)	No
12	Lam & Hwang (2022)	Yes
13	Liao (2022)	Yes
14	Mayol (2018)	Yes
15	Medina Fetterman et al. (2022) Exp.1	No
16	Medina Fetterman et al. (2022) Exp.2	No
17	Patterson et al. (2022)	Yes
18	Rohde & Kehler (2014)	Yes
19	Solstad & Bott (2022) Exp.1	Yes
20	Solstad & Bott (2022) Exp.2	Yes
21	Weatherford & Arnold (2021) Exp.1	Yes
22	Weatherford & Arnold (2021) Exp.2	Yes
23	Zerkle & Arnold (2019)	No
24	Zhan et al. (2020)	Yes
25	Kehler & Rohde (2019)	Yes
26	Portele & Bader (2020)	No

TABLE 7. Dataset availability

Parameter	Prior Type	Estimate	Estimated Error	95% CI
Intercept	Weakly informative prior	2.01	0.63	[0.82, 3.28]
	brms default prior	2.35	0.89	[0.76, 4.28]
VerbType NP1-biased	Weakly informative prior	-0.07	0.62	[-1.28, 1.13]
	brms default prior	-0.07	0.80	[-1.74, 1.43]

TABLE 8. Comparison of model fits using two different priors for analyzing pronoun production referring to predictable versus less predictable subject referents in Rohde & Kehler 2014. Each row presents results for a parameter with both the weakly informative prior and the *brms* default prior.

	Estimate	Estimated error	95% CI
Intercept	-85.00	122.69	[-425.70, -6.42]
Connective so	61.95	117.14	[-22.69, 381.12]

TABLE 9. Model fit using the *brms* default prior comparing the null pronoun production for more predictable object referents (subject-biased ICV + so) vs. null pronoun produced referring to less predictable object referents (subject-biased ICV + because) in Hwang et al. 2022.

Experiment	Year	Language	Pronoun Type	Estimate	Error	95% CIs
Arnold	2001	English	pronoun/null	0.18	0.57	[-1.00, 1.29]
Contemori and Di Domenico (Exp.3)	2021	Italian	pronoun/null	-0.05	0.96	[-1.95, 1.72]
Contemori and Di Domenico (Exp.3)	2021	Italian	overt	-0.34	0.99	[-2.21, 1.73]
Contemori and Di Domenico (Exp.3)	2021	Spanish	pronoun/null	-0.33	0.72	[-1.81, 1.00]
Contemori and Di Domenico (Exp.3)	2021	Spanish	overt	0.35	0.73	[-1.01, 1.85]
Fukumura and van Gompel (Exp.1)	2010	English	pronoun/null	0.04	0.23	[-0.41, 0.48]
Fukumura and van Gompel (Pre.1)	2010	English	pronoun/null	-0.41	0.54	[-1.51, 0.58]
Holler and Suckow (ICV2 + because/since vs. but/although)	2016	German	pronoun/null	0.35	0.23	[-0.08, 0.80]
Holler and Suckow (ICV + because/since)	2016	German	pronoun/null	0.42	0.24	[-0.07, 0.89]
Holler and Suckow (ICV1 + because/since vs. but/although)	2016	German	pronoun/null	0.08	0.25	[-0.42, 0.58]
Hwang et al. (ICV + because)	2022	Mandarin	pronoun/null	-0.40	1.12	[-2.50, 1.89]
Hwang et al. (ICV + because)	2022	Mandarin	overt	-0.19	0.74	[-1.62, 1.27]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	pronoun/null	-1.89	0.81	[-3.45, -0.27]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	overt	-0.02	0.64	[-1.25, 1.29]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	pronoun/null	0.07	1.12	[-2.05, 2.36]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	overt	0.42	0.76	[-1.04, 1.98]
Hwang et al. (TPV + so)	2022	Mandarin	pronoun/null	0.21	1.04	[-1.77, 2.32]
Hwang et al. (TPV + so)	2022	Mandarin	overt	0.83	0.59	[-0.27, 2.06]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	pronoun/null	0.43	0.78	[-1.12, 1.94]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	overt	0.47	0.40	[-0.32, 1.24]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	pronoun/null	0.76	0.97	[-1.18, 2.64]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	overt	0.67	0.55	[-0.35, 1.81]
Hwang continuity	2022	Korean	pronoun/null	2.53	0.51	[1.51, 3.56]
Hwang marker (Exp.1 ICV)	2022	Korean	pronoun/null	0.55	0.71	[-0.87, 1.91]
Hwang marker (Exp.1 TPV)	2022	Korean	pronoun/null	-0.13	0.81	[-1.76, 1.45]
Hwang marker (Exp.2 ICV)	2022	Korean	pronoun/null	-1.72	0.92	[-3.36, 0.32]
Hwang marker (Exp.2 TPV)	2022	Korean	pronoun/null	0.19	0.93	[-1.54, 2.16]
Konuk and von Heusinger	2021	Turkish	pronoun/null	1.83	0.22	[1.41, 2.26]
Konuk and von Heusinger	2021	Turkish	overt	-1.22	0.45	[-2.10, -0.32]
Lam and Hwang	2022	Mandarin	pronoun/null	0.63	0.49	[-0.37, 1.58]
Lam and Hwang	2022	Mandarin	overt	-0.59	0.49	[-1.61, 0.33]
Liao	2022	English	pronoun/null	-0.07	0.21	[-0.49, 0.34]
Mayol	2018	Catalan	pronoun/null	-0.15	0.77	[-1.71, 1.37]
Mayol	2018	Catalan	overt	-0.10	0.76	[-1.59, 1.41]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	pronoun/null	0.05	0.36	[-0.68, 0.71]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	pronoun/null	0.89	0.38	[0.17, 1.65]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	overt	0.31	0.48	[-0.64, 1.23]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	overt	0.83	0.57	[-0.27, 1.98]
Medina Fetterman et al. (written, different gender)	2022	Spanish	pronoun/null	-0.25	0.26	[-0.77, 0.26]
Medina Fetterman et al. (written, same gender)	2022	Spanish	pronoun/null	0.74	0.25	[0.24, 1.23]
Medina Fetterman et al. (written, different gender)	2022	Spanish	overt	-0.95	0.47	[-1.91, -0.10]
Medina Fetterman et al. (written, same gender)	2022	Spanish	overt	-0.42	0.59	[-1.58, 0.69]
Patterson et al.	2022	German	pronoun/null	-0.64	1.14	[-2.89, 1.63]
Rohde and Kelher	2014	English	pronoun/null	-0.06	0.66	[-1.37, 1.17]
Solstad and Bott (Exp.1, because)	2022	German	pronoun/null	0.77	1.15	[-1.54, 2.93]
Solstad and Bott (Exp.1, so)	2022	German	pronoun/null	1.46	1.05	[-0.78, 3.37]
Solstad and Bott (Exp.2)	2022	German	pronoun/null	0.50	0.82	[-0.92, 2.20]
Weatherford and Arnold (Exp.1)	2021	English	pronoun/null	0.25	0.38	[-0.52, 0.97]
Weatherford and Arnold (Exp.2)	2021	English	pronoun/null	0.37	0.45	[-0.53, 1.23]
Zerkle and Arnold	2019	English	pronoun/null	0.80	0.21	[0.39, 1.21]
Zhan et al.	2020	Mandarin	pronoun/null	-1.46	0.86	[-3.08, 0.33]
Zhan et al.	2020	Mandarin	overt	-0.07	0.50	[-1.11, 0.86]
Kehler and Rohde	2019	English	pronoun/null	-0.58	0.70	[-1.94, 0.85]

Portele and Bader	2020	German	pronoun/null	-0.58	0.47	[-1.53, 0.28]

TABLE 10. Summary of the posterior distribution for each comparison pair of subject referents. The estimate represents the effect size (in log odds) for each comparison pair, estimated based on the data gathered from each experiment.

Experiment	Year	Language	Pronoun Type	Estimate	Error	95% CIs
Arnold	2001	English	pronoun/null	0.71	0.36	[0.02, 1.44]
Contemori and Di Domenico (Exp.2)	2021	Italian	pronoun/null	1.81	0.82	[0.21, 3.41]
Contemori and Di Domenico (Exp.2)	2021	Italian	overt	-1.89	0.81	[-3.45, -0.27]
Contemori and Di Domenico (Exp.1)	2021	Spanish	pronoun/null	-0.64	0.55	[-1.77, 0.39]
Contemori and Di Domenico (Exp.1)	2021	Spanish	overt	0.63	0.55	[-0.40, 1.75]
Fukumura and van Gompel (Exp.1)	2010	English	pronoun/null	-0.05	0.22	[-0.48, 0.36]
Fukumura and van Gompel (Pre.1)	2010	English	pronoun/null	0.94	0.28	[0.40, 1.49]
Holler and Suckow (ICV2 + because/since vs. but/although)	2016	German	pronoun/null	0.12	0.18	[-0.25, 0.48]
Holler and Suckow (ICV + because/since)	2016	German	pronoun/null	0.28	0.21	[-0.13, 0.69]
Holler and Suckow (ICV1 + because/since vs. but/although)	2016	German	pronoun/null	0.13	0.22	[-0.31, 0.56]
Hwang et al. (ICV + because)	2022	Mandarin	pronoun/null	0.37	1.27	[-2.00, 2.99]
Hwang et al. (ICV + because)	2022	Mandarin	overt	0.04	1.10	[-2.03, 2.27]
Hwang et al. (ICV1 + because vs. so)	2022	Mandarin	overt	-0.24	1.03	[-2.23, 1.85]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	pronoun/null	-0.16	1.10	[-2.21, 2.11]
Hwang et al. (ICV2 + because vs. so)	2022	Mandarin	overt	0.71	1.09	[-1.37, 2.92]
Hwang et al. (TPV + so)	2022	Mandarin	pronoun/null	-0.43	1.08	[-2.43, 1.85]
Hwang et al. (TPV + so)	2022	Mandarin	overt	0.64	1.11	[-1.47, 2.86]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	pronoun/null	0.47	0.40	[-0.32, 1.24]
Hwang et al. (TPV1 + because vs. so)	2022	Mandarin	overt	0.18	1.06	[-1.87, 2.32]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	pronoun/null	-0.53	0.92	[-2.33, 1.24]
Hwang et al. (TPV2 + because vs. so)	2022	Mandarin	overt	0.07	0.72	[-1.34, 1.44]
Hwang marker (Exp.1 ICV)	2022	Korean	pronoun/null	0.20	1.05	[-1.87, 2.28]
Hwang marker (Exp.1 TPV)	2022	Korean	pronoun/null	-0.23	0.99	[-2.12, 1.68]
Hwang marker (Exp.2 ICV)	2022	Korean	pronoun/null	-0.91	0.94	[-2.65, 1.09]
Konuk and von Heusinger	2021	Turkish	pronoun/null	0.13	0.29	[-0.43, 0.73]
Konuk and von Heusinger	2021	Turkish	overt	-0.62	0.60	[-1.81, 0.51]
Mayol	2018	Catalan	pronoun/null	0.65	0.73	[-0.72, 2.09]
Mayol	2018	Catalan	overt	0.16	0.59	[-0.99, 1.34]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	pronoun/null	0.02	0.89	[-1.77, 1.78]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	pronoun/null	-0.82	0.82	[-2.46, 0.73]
Medina Fetterman et al. (spoken, different gender)	2022	Spanish	overt	0.12	0.39	[-0.64, 0.91]
Medina Fetterman et al. (spoken, same gender)	2022	Spanish	overt	1.54	0.51	[0.59, 2.61]
Medina Fetterman et al. (written, different gender)	2022	Spanish	pronoun/null	-0.58	0.47	[-1.53, 0.32]
Medina Fetterman et al. (written, same gender)	2022	Spanish	pronoun/null	-0.42	0.50	[-1.40, 0.57]
Medina Fetterman et al. (written, different gender)	2022	Spanish	overt	0.79	0.36	[0.14, 1.52]
Medina Fetterman et al. (written, same gender)	2022	Spanish	overt	0.78	0.43	[-0.03, 1.64]
Patterson et al.	2022	German	pronoun/null	0.49	0.55	[-0.57, 1.56]
Rohde and Kelher	2014	English	pronoun/null	-0.24	0.76	[-1.65, 1.35]
Solstad and Bott (Exp.1, because)	2022	German	pronoun/null	0.52	0.83	[-1.20, 2.05]
Solstad and Bott (Exp.1, so)	2022	German	pronoun/null	0.33	0.63	[-0.99, 1.49]
Solstad and Bott (Exp.2)	2022	German	pronoun/null	0.14	0.47	[-0.77, 1.09]
Weatherford and Arnold (Exp.1)	2021	English	pronoun/null	1.29	0.43	[0.41, 2.12]
Weatherford and Arnold (Exp.2)	2021	English	pronoun/null	1.26	0.51	[0.28, 2.29]
Zerkle and Arnold	2019	English	pronoun/null	-0.22	0.35	[-0.92, 0.45]
Zhan et al.	2020	Mandarin	pronoun/null	-0.21	1.12	[-2.35, 2.06]
Zhan et al.	2020	Mandarin	overt	0.83	0.53	[-0.16, 1.92]
Kehler and Rohde	2019	English	pronoun/null	0.83	0.88	[-0.87, 2.63]
Portele and Bader	2020	German	pronoun/null	0.91	0.43	[0.12, 1.77]

TABLE 11. Summary of the posterior distribution for each comparison pair of object referents. The estimate represents the effect size (in log odds) for each comparison pair, estimated based on the data gathered from each experiment.

Model	$ELPD_{\triangle}$	ELPD	$SE_{\triangle}$
all predictors	0	-85.4	0
language family*gram_func_ant, manipulation type	-7.3	-92.7	6.4
language family, gram_func_ant*manipulation type	-15.1	-100.6	13.5
gram_func_ant	-15.7	-101.1	12.0
language family, gram_func_ant	-17.2	-102.6	11.6
language family	-18.1	-103.5	13.8
language family, manipulation type	-18.5	-104.0	13.3
language family, gram_func_ant, manipulation type	-18.7	-104.1	11.0
gram_func_ant, manipulation type	-18.9	-104.3	12.0
manipulation type	-19.2	-104.6	14.0

TABLE 12. Ranking of models in terms of expected log-predictive densities (ELPD, higher is better). $ELPD_{\triangle}$ is the estimated difference in ELPDs from the highest ranked model, and $SE_{\triangle}$ is the standard error of that difference.

Model	$ELPD_{\triangle}$	ELPD	$SE_{\triangle}$
all predictors	0	-108.8	0
pronoun type, manipulation type, language family*gram_func_ant, pronoun type:gram_func_ant	-1.3	-110.1	3.3
pronoun type*manipulation type, language family*gram_func_ant, pronoun type:gram_func_ant, gram_func_ant:manipulation type	-1.6	-110.4	4.1
pronoun type*manipulation, language family*gram_func_ant, gram_func_ant:manipulation type	-2.1	-110.9	5.5
pronoun type, manipulation type, language family*gram_func_ant, gram_func_ant:manipulation type	-2.2	-110.9	4.5
language family*gram_func_ant, pronoun type*manipulation type	-2.3	-111.1	6.6
pronoun type, manipulation type, language family*gram_func_ant	-2.4	-111.2	5.4
manipulation, language family, pronoun type*gram_func_ant	-3.0	-111.8	8.5
language family, manipulation type, pronoun type*gram_func_ant, gram_func_ant:manipulation type	-4.9	-113.7	9.2
gram_func_ant, manipulation type	-6.6	-115.4	12.8
manipulation type, language family	-7.0	-115.8	14.1
gram_func_ant, manipulation type, language family	-7.0	-115.8	11.9
gram_func_ant, language family, pronoun type*manipulation type	-7.7	-116.5	12.1
gram_func_ant, language family	-7.8	-116.6	11.5
manipulation	-7.9	-116.7	15.8
gram_func_ant, pronoun type, manipulation type	-8.1	-116.9	13.2
pronoun type, manipulation type	-8.3	-117.1	15.1
gram_func_ant, pronoun type, language family	-8.4	-117.2	10.7
pronoun type, manipulation type, language family	-8.8	-117.6	14.5
gram_func_ant	-9.5	-118.3	12.8
language family	-9.5	-118.3	14.8
gram_func_ant, pronoun type	-9.6	-118.4	11.3
gram_func_ant, pronoun type, manipulation type, language family	-9.7	-118.4	12.2
pronoun type, language family	-10.2	-119.0	14.6
gram_func_ant*manipulation type, pronoun type, language family	-10.7	-119.5	11.7
pronoun type, language family, gram_func_ant*manipulation type, pronoun type:manipulation type	-11.3	-120.1	12.1
pronoun type	-12.2	-121.0	15.2

TABLE 13. Ranking of models in terms of expected log-predictive densities (ELPD, higher is better). $ELPD_{\triangle}$ is the estimated difference in ELPDs from the highest ranked model, and $SE_{\triangle}$ is the standard error of that difference.

Model	Estimate	Est. Err	95% CI	BF
all predictors	0.51	0.45	[-0.31, 1.46]	9.15
pronoun type, manipulation type, language family*gram_func_ant, pronoun type:gram_func_ant	0.48	0.27	[-0.02, 1.01]	10.2
pronoun type*manipulation type, language family*gram_func_ant, pronoun type:gram_func_ant, gram_func_ant:manipulation type	0.35	0.23	[-0.08, 0.85]	18.51
pronoun type*manipulation, language family*gram_func_ant, gram_func_ant:manipulation type	0.36	0.24	[-0.09, 0.84]	17.02
pronoun type, manipulation type, language family*gram_func_ant, gram_func_ant:manipulation type	0.40	0.25	[-0.11, 0.89]	19.51
language family*gram_func_ant, pronoun type*manipulation type	0.46	0.22	[0.05, 0.93]	59.61
pronoun type, manipulation type, language family*gram_func_ant	0.48	0.28	[-0.09, 1.04]	23.1
manipulation, language family, pronoun type*gram_func_ant	0.46	0.26	[-0.05, 0.99]	29.08
language family, manipulation type, pronoun type*gram_func_ant, gram_func_ant:manipulation type	0.44	0.25	[-0.04, 0.97]	30.75
gram_func_ant, manipulation type	0.34	0.20	[-0.06, 0.74]	23.24
manipulation type, language family	0.47	0.27	[-0.03, 0.99]	27.99
gram_func_ant, manipulation type, language family	0.50	0.25	[0.01, 1.02]	41.11
gram_func_ant, language family, pronoun type*manipulation type	0.46	0.21	[0.06, 0.90]	61.5
gram_func_ant, language family	0.34	0.26	[-0.18, 0.87]	12.65
manipulation	0.32	0.20	[-0.06, 0.73]	20.86
gram_func_ant, pronoun type, manipulation type	0.34	0.20	[-0.06, 0.76]	22.95
pronoun type, manipulation type	0.30	0.20	[-0.08, 0.71]	18.7
gram_func_ant, pronoun type, language family	0.33	0.25	[-0.17, 0.82]	11.9
pronoun type, manipulation type, language family	0.46	0.24	[-0.02, 0.94]	32.9
gram_func_ant	0.22	0.19	[-0.17, 0.62]	7.73
language family	0.31	0.27	[-0.24, 0.82]	10.11
gram_func_ant, pronoun type	0.21	0.18	[-0.17, 0.59]	7.77
gram_func_ant, pronoun type, manipulation type, language family	0.50	0.24	[0.05, 1.00]	56.14
pronoun type, language family	0.31	0.25	[-0.22, 0.84]	10.2
gram_func_ant*manipulation type, pronoun type, language family	0.50	0.27	[-0.03, 1.05]	31
pronoun type, language family, gram_func_ant*manipulation type, pronoun type:manipulation type	0.46	0.21	[0.07, 0.93]	56.97
pronoun type	0.20	0.19	[-0.17, 0.59]	6.68

TABLE 14. Intercept estimates and Bayes Factor values for all models evaluated in the sensitivity analysis.

Predictor	level	Count
Language family	Turkic	2
	Mandarin	14
	Korean	8
	Germanic	33
	Romance	14
Manipulation type	ICV	35
	relativeClause	2
	relation	19
	TPV	15
Grammatical function of the antecedent	object	33
	subject	38

TABLE 15. Number of datapoints by predictor and level for the analysis in Section 4.2.

Predictor	Level	Count
Language family	Turkic	4
	Mandarin	29
	Korean	8
	Romance	28
Manipulation type	ICV	28
	relation	20
	TPV	21
Grammatical function of the antecedent	object	32
	subject	37
Pronoun type	overt	31
	null	38

TABLE 16. Number of datapoints by predictor and level for the analysis in Section 4.3.

		Manipulation Type			
Language Family	Grammatical Function	ICV	Relative Clause	Relation	TPV
Turkic	Object	1	0	0	0
Mandarin	Object	2	0	3	1
Korean	Object	1	0	2	0
Germanic	Object	10	1	3	2
Romance	Object	3	0	0	4
Turkic	Subject	1	0	0	0
Mandarin	Subject	3	0	4	1
Korean	Subject	1	0	3	1
Germanic	Subject	10	1	4	2
Romance	Subject	3	0	0	4

TABLE 17. Number of datapoints by language family, grammatical function, and manipulation type in Section 4.2.

	Estimated Mean	Estimated Error	95% CI
Intercept	0.23	0.19	[-0.16, 0.60]
gram_func_ant object	0.11	0.11	[-0.11, 0.33]
manipulationType ICV	0.05	0.20	[-0.37, 0.44]
manipulationType relativeClause	-0.17	0.46	[-1.12, 0.74]
manipulationType relation	-0.13	0.21	[-0.55, 0.30]
gram_func_ant object:manipulationType ICV	-0.04	0.13	[-0.29, 0.21]
gram_func_ant object:manipulationType relativeClause	0.54	0.31	[-0.06, 1.15]
gram_func_ant object:manipulationType relation	-0.07	0.13	[-0.33, 0.19]

TABLE 18. Effect estimate (in log odds) of predictability on the use of pronouns in Germanic languages: Summary of the model with two predictors (manipulation type and grammatical function of the antecedent), including interaction terms.

	Estimate	Estimated error	95% CI
Intercept	0.43	0.36	[-0.28, 1.17]
manipulationType ICV	0.01	0.28	[-0.56, 0.56]
manipulationType relativeClause	-0.15	0.69	[-1.49, 1.24]
manipulationType relation	-0.03	0.28	[-0.58, 0.52]
languageFamily Turkish	0.54	0.67	[-0.83, 1.84]
languageFamily Mandarin	-0.62	0.49	[-1.61, 0.34]
languageFamily Korean	0.14	0.55	[-0.91, 1.27]
languageFamily Germanic	-0.22	0.33	[-0.90, 0.43]
gram_func_ant object	-0.03	0.15	[-0.33, 0.28]
languageFamily Turkish:gram_func_ant object	-0.83	0.18	[-1.18, -0.47]
languageFamily Mandarin:gram_func_ant object	0.09	0.23	[-0.35, 0.55]
languageFamily Korean:gram_func_ant object	-0.00	0.27	[-0.54, 0.53]
languageFamily Germanic:gram_func_ant object	0.14	0.12	[-0.10, 0.37]
gram_func_ant object:manipulationType ICV	0.00	0.13	[-0.25, 0.26]
gram_func_ant object:manipulationType relativeClause	0.55	0.32	[-0.06, 1.20]
gram_func_ant object:manipulationType relation	-0.12	0.13	[-0.38, 0.14]

TABLE 19. Effect estimate (in log odds) of predictability on the use of the most reduced referential form: Summary of the model with three covariates (manipulation type, language family, and grammatical function of the antecedent) and interactions between them, with the Spanish data excluded.

	Estimate	Estimated error	95% CI
Intercept	0.42	0.62	[-0.82, 1.68]
gram_func_ant object	-0.36	0.19	[-0.73, 0.00]
pronounType overt	-0.29	0.11	[-0.50, -0.07]
manipulationType ICV	-0.33	0.29	[-0.91, 0.25]
manipulationType relation	0.09	0.23	[-0.36, 0.52]
languageFamily Turkic	0.35	1.32	[-2.29, 3.18]
languageFamily Mandarin	-0.03	0.92	[-1.95, 1.73]
languageFamily Korean	-0.18	1.01	[-2.23, 1.98]
gram_func_ant object:pronounType overt	0.05	0.10	[-0.15, 0.26]
gram_func_ant object:manipulationType ICV	0.48	0.21	[0.07, 0.89]
gram_func_ant object:manipulationType relation	-0.15	0.19	[-0.53, 0.22]
gram_func_ant object:languageFamily Turkic	-0.73	0.19	[-1.10, -0.36]
gram_func_ant object:languageFamily Mandarin	0.30	0.19	[-0.08, 0.68]
gram_func_ant object:languageFamily Korean	0.36	0.27	[-0.17, 0.90]

TABLE 20. Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with four covariates and their interactions added (manipulation type, language family, grammatical function of the antecedent, and pronoun type), with Spanish data excluded.

	Estimate	Estimated Error	95% CI
Intercept	0.32	0.34	[-0.35, 1.02]
manipulationType ICV	-0.00	0.27	[-0.53, 0.53]
manipulationType relativeClause	-0.13	0.70	[-1.53, 1.32]
manipulationType relation	-0.02	0.28	[-0.62, 0.52]
languageFamily Turkish	0.64	0.67	[-0.69, 1.99]
languageFamily Mandarin	-0.49	0.49	[-1.48, 0.45]
languageFamily Korean	0.27	0.57	[-0.79, 1.43]
languageFamily Germanic	-0.15	0.34	[-0.84, 0.51]
gram_func_ant object	-0.11	0.14	[-0.39, 0.16]
languageFamily Turkish:gram_func_ant object	-0.78	0.18	[-1.13, -0.43]
languageFamily Mandarin:gram_func_ant object	0.17	0.22	[-0.27, 0.59]
languageFamily Korean:gram_func_ant object	0.08	0.26	[-0.45, 0.59]
languageFamily Germanic:gram_func_ant object	0.22	0.11	[0.02, 0.43]
gram_func_ant object:manipulationType ICV	0.03	0.13	[-0.22, 0.29]
gram_func_ant object:manipulationType relativeClause	0.54	0.32	[-0.10, 1.16]
gram_func_ant object:manipulationType relation	-0.12	0.13	[-0.38, 0.15]

TABLE 21. Effect estimate (in log odds) of predictability on the use of the most reduced reference form: summary of the model with three covariates and their interactions added (manipulation type, language family, and grammatical function of the antecedent), with Contemori & Di Domenico 2021 and Solstad & Bott 2022 excluded.

	Estimate	Estimated error	95% CI
Intercept	0.48	0.54	[-0.59, 1.61]
gram_func_ant object	-0.17	0.10	[-0.36, 0.03]
pronounType overt	-0.11	0.06	[-0.23, 0.01]
manipulationType ICV	-0.39	0.26	[-0.91, 0.12]
manipulationType relation	0.09	0.22	[-0.36, 0.52]
languageFamily Turkic	0.50	1.15	[-1.78, 2.93]
languageFamily Mandarin	-0.15	0.80	[-1.82, 1.47]
languageFamily Korean	0.08	0.94	[-1.76, 2.09]
gram_func_ant object:pronounType overt	0.17	0.07	[0.04, 0.30]
gram_func_ant object:manipulationType ICV	0.33	0.14	[0.07, 0.59]
gram_func_ant object:manipulationType relation	-0.21	0.18	[-0.56, 0.14]
gram_func_ant object:languageFamily Turkic	-0.70	0.19	[-1.06, -0.34]
gram_func_ant object:languageFamily Mandarin	0.13	0.17	[-0.19, 0.46]
gram_func_ant object:languageFamily Korean	0.36	0.26	[-0.16, 0.89]

TABLE 22. Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with four covariates and their interactions added (manipulation type, language family, grammatical function of the antecedent, and pronoun type), with Contemori & Di Domenico 2021 and Solstad & Bott 2022 excluded.

	Estimate	Estimated error	95% CI
Intercept	0.34	0.34	[-0.33, 1.03]
manipulationType ICV	0.01	0.29	[-0.56, 0.57]
manipulationType relativeClause	-0.19	0.74	[-1.63, 1.25]
manipulationType relation	-0.02	0.29	[-0.59, 0.55]
language German	-0.16	0.45	[-1.07, 0.71]
language Mandarin	-0.56	0.55	[-1.65, 0.50]
language Catalan	-0.11	0.86	[-1.79, 1.57]
language Spanish	-0.62	0.54	[-1.71, 0.47]
language Turkish	0.65	0.75	[-0.83, 2.13]
language Korean	0.27	0.61	[-0.86, 1.56]
language English	-0.08	0.39	[-0.85, 0.71]
gram_func_ant object	0.08	0.15	[-0.22, 0.38]
language German:gram_func_ant object	-0.08	0.14	[-0.36, 0.19]
language Mandarin:gram_func_ant object	-0.07	0.24	[-0.54, 0.41]
language Catalan:gram_func_ant object	0.26	0.40	[-0.52, 1.04]
language Spanish:gram_func_ant object	0.09	0.19	[-0.28, 0.46]
language Turkish:gram_func_ant object	-0.94	0.19	[-1.32, -0.57]
language Korean:gram_func_ant object	-0.16	0.30	[-0.75, 0.44]
language English:gram_func_ant object	0.09	0.14	[-0.17, 0.36]
manipulationType ICV:gram_func_ant object	0.00	0.13	[-0.25, 0.25]
manipulationType relativeClause:gram_func_ant object	0.49	0.32	[-0.15, 1.11]
manipulationType relation:gram_func_ant object	-0.02	0.16	[-0.34, 0.30]

TABLE 23. Effect estimate (in log odds) of predictability on the use of the most reduced reference form: Summary of the model with manipulation type, language, grammatical function of the antecedent and their interactions added as covariates.

	Estimate	Estimated error	95% CI
Intercept	0.43	0.55	[-0.69, 1.53]
gram_func_ant object	-0.15	0.12	[-0.39, 0.10]
pronounType overt	-0.11	0.06	[-0.23, 0.02]
manipulationType ICV	-0.43	0.26	[-0.93, 0.06]
manipulationType relation	0.08	0.22	[-0.34, 0.50]
language Mandarin	0.00	0.85	[-1.69, 1.84]
language Catalan	0.13	1.30	[-2.48, 2.82]
language Spanish	-0.46	0.85	[-2.34, 1.23]
language Turkish	0.55	1.34	[-2.19, 3.12]
language Korean	0.17	1.05	[-1.82, 2.28]
gram_func_ant object:pronounType overt	0.15	0.06	[0.02, 0.27]
gram_func_ant object:manipulationType ICV	0.26	0.14	[-0.00, 0.54]
gram_func_ant object:manipulationType relation	-0.24	0.18	[-0.59, 0.11]
gram_func_ant object:language Mandarin	0.15	0.19	[-0.22, 0.51]
gram_func_ant object:language Catalan	0.12	0.28	[-0.44, 0.70]
gram_func_ant object:language Spanish	0.09	0.22	[-0.34, 0.51]
gram_func_ant object:language Turkish	-0.67	0.19	[-1.05, -0.30]
gram_func_ant object:language Korean	0.34	0.29	[-0.22, 0.91]

TABLE 24. Effect estimate (in log odds) of predictability in null-subject languages: summary of the model with manipulation type, language, grammatical function of the antecedent, pronoun type and their interactions added as covariates.

	Estimate	Estimated Error	95% CI
Intercept	0.24	0.16	[-0.09, 0.57]
modality written	-0.13	0.14	[-0.42, 0.13]

TABLE 25. Effect estimate (in log odds) of predictability on the use of the most reduced reference form in English and Spanish: Summary of the model with task modality added as the only covariate.

	Estimate	Estimated Error	95% CI
Intercept	0.28	0.16	[-0.05, 0.60]
gender different	-0.09	0.09	[-0.27, 0.09]

TABLE 26. Effect estimate (in log odds) of predictability on the use of the most reduced reference form in non-corpus studies: Summary of the model with character gender added as the only covariate.

	Estimate	Estimated Error	95% CI
Intercept	0.31	0.16	[0.00, 0.62]
paradigm free	-0.13	0.15	[-0.41, 0.16]

TABLE 27. Effect estimate (in log odds) of predictability on the use of the most reduced reference form: Summary of the model with completion paradigm added as the only covariate.