Collecting data in understudied language varieties: A methodological note

Laia Colina Fortuny[*], Roberta D'Alessandro[*], and M. Carmen Parafita Couto[#]
[*]*Utrecht University*, [#]*University of Santiago de Compostela*

Abstract

Eliciting linguistic data from speakers of non-standardized varieties presents well-known challenges. Traditional acceptability judgment tasks have been widely used to probe grammatical knowledge, but task design, such as mode of presentation, response options, and labeling, can strongly influence outcomes. In this study, we investigate a small group of heritage Catalan speakers in Germany and the Netherlands using a repetition-based task rather than a standard judgment task. Our findings indicate that responses to repetition tasks differ significantly from commonly known generalizations on the phenomenon at issue investigated through grammaticality judgment tasks. We conclude that repetition tasks can uncover patterns of grammatical representation that may differ from those captured by judgment tasks, and that they provide deeper insights into grammatical representation than the latter. This suggests that carefully designed elicitation tasks can mitigate common biases and more accurately reveal the grammatical representations of heritage speakers. These findings highlight both the challenges and the potential of alternative methodologies for studying underrepresented languages.

1. Introduction

Collecting data for non-standardized varieties is a challenge for linguists. Data elicitation in the form of acceptability judgments has largely been proved to be effective for English monolingual speakers, most notably by Sprouse et al. (2013) and Sprouse and Almeida (2017). That this might not be the whole story has, however, been noted by Marty et al. (2020), who showed that the task configuration can influence acceptability judgment tasks. This, in turn, means that the replication of the judgments can be affected by the way in which the tasks are designed. Specifically, Marty et al. identify three factors that can influence acceptability judgments: the mode of presentation, the number of response options, and the use of labels. The mode of presentation refers to whether stimuli are presented visually, aurally, or in written form, which can affect comprehension and attention. The number of response options, for example, binary "acceptable/ unacceptable" choices versus Likert scales, can influence how speakers categorize linguistic phenomena, potentially masking subtle gradations in acceptability. Finally, the use of labels concerns whether linguistic terms (e.g., "grammatical", "acceptable") are provided to guide participants' responses; this can alter participants' interpretations of the task and introduce response biases. Taken together, these factors highlight that acceptability judgments are sensitive not only to the linguistic knowledge of participants but also to the design of the task itself.

In this paper, we show further evidence that different tasks might result in slightly different results. Although languages with very few speakers can be problematic for establishing generalizations (see also Leivada et al. 2019, D'Alessandro et al. 2021, and Andriani et al. 2022 on this issue), we will also show that such generalizations can in fact be drawn if appropriate procedural care is applied.

Our empirical base is a small group of heritage Catalan speakers in Germany and the Netherlands. We employ a repetition-based task that indirectly taps into speakers' grammatical knowledge, showing that this approach can produce different results from acceptability judgment tasks and reveal aspects of grammatical representation that might not emerge in explicit judgments. We show that participants' responses are more closely aligned with their underlying grammatical knowledge when the task is formulated to directly engage with these representations. Unlike explicit judgments, which allow for post-interpretive reflection and social bias, the repetition task functions as a reconstructive process. To repeat a sentence accurately, particularly under a cognitive load, speakers must process the stimulus through their own internal grammar before regenerating it. This method provides a window into the speaker's implicit processing, similar to the real-time sensitivity observed in online tasks like self-paced reading (Tokaç-Scheffer et al. 2023). By using an auditory repetition format, we mitigate the literacy biases and "yes-bias"

 Nordlyd 49.1: 3–16, Special issue in honor of Johan Rooryck, edited by Jeroen van Craenenbroeck, Aysa Ekanger, Camil Staps, and Guido Vanden Wyngaerd. Published at the University of Tromsø – The Arctic University of Norway.
http://septentrio.uit.no/index.php/nordlyd https://doi.org/10.7557/12.8324

common in written acceptability judgment tasks, allowing for a more accurate mapping of the speaker's active grammatical boundaries.

2. The yes-bias

Bilingual respondents routinely display a "yes-bias" (Polinsky 2006) in judgment tasks: a tendency to accept sentences as grammatical or "OK", even when their internal competence would suggest otherwise. The yes-bias, i.e. the "reluctance to reject ungrammatical material" (Polinsky 2018:95) was first described for L2 speakers by Ellis (1991) and observed to be at work also for heritage speakers by Polinsky (2006) in an article on heritage Russian speakers in America. Romano and Guijarro Fuentes (2023) further note that this effect is magnified in metalinguistic judgment tasks for heritage grammars, where practical communication norms are overruled by formal prescriptive standards. The yes-bias is caused by multiple factors:

(i) Deference to the perceived authority of the researcher or stimulus provider.

(ii) Socialized tolerance of variation in bilingual communities.

(iii) Uncertainty about minority-language norms, especially when the heritage language is primarily spoken at home or in informal settings.

For heritage speakers, explicit ratings may thus underrepresent the boundaries of their active grammatical knowledge, and contribute to the "incomplete acquisition" narrative (Polinsky 2006; 2008, Montrul 2002; 2008, but see Putnam and Sánchez 2013) that fails to capture their dynamic competence (cf. Leivada et al. 2023).

Despite variable proficiency, heritage speakers have been shown to retain systematic representations of morphosyntax (Scontras et al. 2018). However, explicit judgments frequently conflict with production and comprehension data; heritage speakers may "accept" ungrammatical forms out of politeness or uncertainty, without producing such forms in spontaneous speech themselves. Studying implicit representations requires methods less reliant on conscious reflection or explicit evaluation. The potential for explicit ratings to underrepresent the grammatical knowledge of heritage speakers is rooted in the high metalinguistic demands of such tasks, which often require speakers to perform conscious evaluations that may not reflect their true underlying competence. As noted by Van Baal (2025), acceptability judgments can be particularly challenging for speakers of heritage or moribund varieties due to yes-bias, prescriptive pressures, or a lack of confidence in formal registers. When heritage speakers appear to accept ungrammatical structures or show high variability in offline ratings, it often fuels the "incomplete acquisition" narrative (Montrul 2008, Polinsky 2018), which interprets these results as a lack of stable mental representations. However, time-sensitive or implicit measures frequently tell a different story. For instance, Tokaç-Scheffer et al. (2023) demonstrated that even when heritage speakers struggle with explicit judgments, their online processing (measured via self-paced reading) reveals real-time sensitivity to grammatical features like evidentiality. By bypassing the "metalinguistic filter" of explicit ratings, alternative methods, such as the repetition task used in this study, can reveal a dynamic competence that remains systematic and robust, even when it does not align with the prescriptive norms measured by standard judgment tasks. This is especially important for phenomena like agreement, where intuitions can be subtle and context-sensitive. To illustrate this point, we present a small experiment on gender in heritage speakers.

2.1. Gender and number agreement in heritage speakers

Gender and number attribution and agreement are widely considered in heritage language studies (see Polinsky 2008, Albirini et al. 2013, Rodríguez and Reglero 2015, Lohndal and Westergaard 2016; 2021, Jegerski and Fernandez Cuenca 2025, Busterud et al. 2025, and others). Our study had the aim of checking whether the generalizations that were found through different kinds of tests, mainly grammaticality judgment tasks, could be replicated by our population. Before going into the details of our study, a terminological note is in order.

The term *agreement* refers to at least three different phenomena in linguistics: the first one is the attribution of lexical gender or number to an item. We will refer to that as "morphological agreement". It is usually assumed (much depending on one's theory of the lexicon) that this kind of gender is assigned directly to the lexical item. For instance, the word for 'chair' in Catalan is *cadira*. This lexical item is feminine singular, and this gender is independent of any agreement relation that this word might entertain with other elements in the clause. Especially in languages like Catalan, which are heavily morphologically gender-marked, gender and number are also visible on all the items occurring within an NP. We refer to that as "NP-internal agreement" or "Concord". For instance, 'the old and beautiful chair' in Spanish is *la cadir**a** vell**a** i bonic**a***, where the feminine singular morpheme *-a* appears on all items within the NP. This phenomenon is similar to "spreading" the ending through the phrase. The third kind of agreement is sentential, of which there are several subtypes. We will not consider agreement that occurs between the subject and the verb in this paper as it is not directly related to gender. A second type of sentential agreement is the sort of agreement that occurs in predicative sentences, for instance, or in secondary predication. The Catalan equivalent of English "Mary considers the chair too old" would be *La Maria considera la cadira massa vella,* where *vella* will be agreeing with *la cadira* in the same way as it would in a simple predicative structure like *La cadira és vella*, translated as "The chair is old". Catalan also has participial agreement, though it is more constrained and less consistently realized than, for example, Italian, so sentential agreement is mostly limited to the subject with the auxiliary, which does not show gender.

In this paper, we will be considering sentential agreement in predicative contexts with the aim of ascertaining whether heritage speakers have stable grammatical representations for long-distance syntactic sentential agreement, or whether their patterns reflect variability when compared to baseline Catalan speakers. Moving away from deficiency-oriented views of heritage speakers, which have been extensively criticized in the literature (Putnam and Sánchez 2013, Kupisch and Rothman 2016, among others), we adopt a non-deficit perspective on heritage grammars by treating differences from baseline varieties as instances of grammatical variability rather than grammatical incompleteness. Moreover, we challenge the problematic comparisons between heritage speakers and monolingual speakers (Rothman et al. 2023) by instead comparing bilinguals with bilinguals, which ensures a more ecologically valid comparison group.

2.1.1. Long-distance agreement and agreement attraction

Agreement and agreement attraction, i.e. the phenomenon whereby speakers accept agreement with the linearly closest intervening element, have been widely studied, both for monolingual speakers and for heritage speakers. Given the limited literature on Catalan as a heritage language, we take as our point of departure a study on Spanish, under the assumption that its findings can be extended to Catalan because of the almost complete overlap in agreement patterns between the two languages. The study is by Scontras et al. (2018), who investigate gender and number agreement in heritage Spanish speakers in the US. Their intent was different from ours: they were trying to check whether features probe separately, and whether principles of featural economy are at work, reducing agreement and bundling together features that normally agree separately. What matters for us is that they found that heritage Spanish speakers present strong deviations with respect to monolinguals when it comes to acceptability judgments. Before we examine Scontras et al.'s results, it is important to underline that the contact language they considered was English, and not German or Dutch. The contact language, as well as the language environment, can also play a role in the output of grammatical change (see for instance the study by Özsoy et al. 2022 on Turkish varieties in different language ecologies). For this short paper, we will set aside these small differences, especially given that both Dutch and German have rather reduced NP-internal agreement and therefore their effect can be considered equivalent to that of English. However, more work needs to be done in this respect.

Scontras et al. (2018) performed an auditory rating task: speakers were asked to rate the acceptability of a grammar on a 1–5 Likert scale. The results are very interesting: grammaticality effects are found only in the feminine, meaning that heritage speakers accept agreement of a feminine with a (default) masculine, but not vice versa. More than the results themselves, what is particularly interesting is that these effects are exactly the same as those found in monolingual speakers. In other words, no difference has been observed in the acceptability of long-distance agreement sentences for gender between monolinguals and heritage

speakers: what is different is that monolinguals consider the ungrammaticality of each feature (number or gender) separately, and therefore are also sensitive to the single featural mismatch, while heritage speakers cluster the two dimensions together.

For what concerns us, the interesting results are that in sentences with a long-distance agreement relationship heritage speakers accept ungrammatical agreement attraction by a linearly closer intervener, though it is also worthwhile considering that monolingual speakers also have the same effect (with respect to gender, not clustering). Scontras et al. reflect on the possibility that this acceptance might not stem only from a yes-bias but also from the heavy computational load that long-distance agreement imposes on speakers.

After considering these results, we set ourselves to check whether they would be the same if a different test were used. Like Scontras et al., we provided auditory stimuli. In order to check whether the speakers' judgments were informed by the yes-bias, by syntactic complexity, or by the task, we decided to perform a production task in the form of repetition and correction: our informants were instructed to correct the sentences that they found ungrammatical, and repeat them correctly. If the problem were only in the length of the sentences, practically equivalent to those in Scontras et al. (2018), we would expect the same results. If different tests yield different results, this would imply that we need to be cautious about the generalizations derived from any elicitation.

Before moving to the description of the test, one disclaimer is in order: our study was a small methodological exercise conducted under severe time constraints. This means that our findings should be regarded as a starting point for further investigation rather than as a solid generalization.

3. The test

3.1. Research questions and hypotheses

The study examined whether elicited imitation tasks could tap into heritage speakers' implicit knowledge of gender and number agreement in Catalan, and whether this method would reveal error patterns different from those observed in judgment tasks. The main predictions were:

1. Heritage speakers would "repair" ungrammatical agreement more often in repetition than acknowledge violations in judgment tasks.
2. The error profiles for number versus gender agreement attraction would help clarify which feature is more vulnerable in heritage grammars.

3.2. Methodology

3.2.1. Participants

A total of 37 informants participated in the present study. They constitute three main groups: Catalan heritage speakers, first-generation Catalan speakers and mainland Catalan speakers. The two immigrant groups (heritage speakers and first-generation speakers) are further subdivided by country of residence: Germany and the Netherlands, thus forming four subgroups. In the larger study for which this data was collected, mainland speakers were also divided between two different generations: a younger generation similar in age to the heritage speakers, and an older generation similar in age to the first-generation speakers. However, the present study does not investigate age-related variation within mainland speakers, and they are therefore treated as a single group. In this section, these subdivisions are described separately to better characterize the participants, but for the purpose of the analysis, we will maintain the division into three main groups.

Heritage speakers, as mentioned, are divided into two subgroups. The German group consisted of 6 speakers (2 male; 3 female; 1 non-binary/other), aged between 21 and 38 (mean age = 28.3; standard deviation [SD] = 6.5), and the Dutch group consisted of 3 speakers (1 male; 2 female), aged between 28 and 43 (mean age = 36.3; SD = 7.6). In terms of linguistic profile, the groups were highly homogeneous. All grew up in households with at least one Catalan parent and were exposed to both the majority language

	Heritage speakers Germany (n = 6)		Heritage speakers Netherlands (n = 3)		First-generation Germany (n = 10)		First-generation Netherlands (n = 10)	
	German	Catalan	Dutch	Catalan	German	Catalan	Dutch	Catalan
Speaking	3.0	2.2	3.0	2.0	2.7	3.0	2.0	2.9
Understanding	3.0	2.8	3.0	2.3	2.8	3.0	2.4	3.0
Reading	2.8	2.0	3.0	1.7	2.7	3.0	2.2	3.0
Mean	2.9	2.3	3.0	2.0	2.7	3.0	2.2	3.0

Table 1: Mean of self-reported ratings of Catalan, German, and Dutch proficiency by speaker group.

(German or Dutch) and Catalan from birth. The only exception was a US-born participant from the German group: although born to two Catalan parents, their exposure to German began at age 2, after moving to Germany. Moreover, all parents originated from the same Catalan region, namely, the province of Barcelona. While three participants had a year of formal Catalan instruction, the entire group was educated in the local majority language and self-identified as being more proficient in German or Dutch than in Catalan.

Language exposure and use were evaluated through two distinct parameters. First, participants identified their five most frequent interlocutors and the languages they used with each of them. The German-based group reported that German dominated their daily interactions (65.6%), with Catalan accounting for only 18.8%, and the remaining 15.6% distributed among English, Spanish, and/or Portuguese.[1] Similarly, the Dutch-based group used Dutch with 75% of their primary contacts, while the remaining 25% of interactions occurred in Catalan. The second parameter involved ranking exposure to German or Dutch and Catalan from 0 'never' to 5 'always' across different contexts (family, friends, reading, TV and series, and radio, music and podcasts). Both heritage groups reported moderate Catalan exposure within the family (mean = 3), but reported significant low levels of exposure in other domains (mean < 2 for Germany; mean < 1 for the Netherlands). Contrarily, exposure to the majority languages was high, especially with friends (mean = 4.17 in Germany; mean = 4.7 in the Netherlands). These metrics confirm a clear dominance in German or Dutch over Catalan. This imbalance is further reflected in participants' self-rated proficiency (scale from 1 'low' to 3 'high') across speaking, listening, and reading. A summary of the results, alongside first-generation speakers' data, is shown in Table 1. As can be observed, both groups reported higher overall competence in the majority language than in their heritage language.

The first-generation immigrant group consists of 20 Catalan speakers: 10 living in Germany (10 female), aged 28–63 (mean age = 46.5; SD = 9.7), and 10 living in the Netherlands (2 male; 8 female), aged 33–63 (mean age = 49.0; SD = 9.9). All emigrated from Catalonia and, in one instance, from the Balearic Islands, between the ages of 18 and 34, and all have resided in Germany or the Netherlands for at least 10 years (Germany: mean length of residence = 19.5; SD = 7.3; Netherlands: mean length of residence = 17.2; SD = 7.0). Geographically, most are from the province of Barcelona, matching the heritage speakers' parents. Only two are from Girona and one is from the Balearic Islands. With the exception of the two oldest participants, all were educated in Catalan; the outliers were schooled during or immediately after the Spanish Francoist dictatorship (1939–1975), a period when education in Catalan was prohibited. Regarding language dominance, 18 participants identified Catalan as their most fluent language, one cited German, and the remaining one English, followed by Dutch.

The evaluation of language exposure and use for first-generation immigrants followed the same parameters as the heritage group. In terms of social interaction, those in Germany reported using German

[1] This suggests a multilingual rather than bilingual background. Although additional languages could potentially influence Catalan attainment and change, analyzing those specific effects is beyond the current study's scope. Consequently, these other languages will be included in the analysis only if the data reveal significant outliers or unexpected deviations.

in 52.5% of daily conversations, Catalan in 27.9%, and Spanish or English in the remaining 19.6%. In contrast, first-generation speakers from the Netherlands used Dutch with only 24.1% of their interlocutors, with Catalan (44.4%) and a combination of Spanish and English (31.5%) making up the rest. Regarding exposure, both groups showed the highest Catalan engagement within the family (mean = 3.7). However, a divergence emerged in other contexts: while the group from Germany reported higher exposure to German than Catalan across all domains, the group from the Netherlands maintained higher Catalan exposure in all contexts except with friends (mean = 1.8). These differences in exposure and use between the two first-generation groups are mirrored in the dominant language proficiency ratings in Table 1, which are higher for German than for Dutch.

The mainland Catalan speakers constitute the control group, which provided an age-matched baseline for the two generations represented in the experimental groups. This group consisted of 8 speakers: a younger subset (n = 4; 2 male, 1 female, 1 non-binary/other) aged 18–25, and an older subset (n = 4; 2 male, 2 female) aged 51–60. All control speakers are from the province of Barcelona, consistent with the heritage speakers' parents and the first-generation group. Catalan is the dominant language for all mainland participants, with only one individual reporting equal dominance in Catalan and Spanish.

First-generation and heritage speakers were recruited through social media and snowball sampling, whereas mainland speakers were reached mainly through the researcher's personal network.

3.2.2. Stimuli

The task consisted of 30 stimuli: 18 experimental items, 4 control sentences, and 8 fillers. Because the task was part of a larger study (Colina Fortuny 2025), we used the filler items to test the phenomenon of interest, namely, sentences that can trigger agreement attraction effects. This specific distribution was thus chosen to provide enough critical stimuli for the agreement attraction analysis while limiting the overall length of the experimental session to prevent participant fatigue.

The stimuli were adapted from the materials of Fuchs et al. (2015), later used in Scontras et al. (2018). With the aim of using a structure where a noun intervenes between the source and target of agreement, and where all agreeing elements are marked for both gender (masculine vs feminine) and number (singular vs plural), they designed sentences with a small clause, with the following syntactic structure:

(1) (Subject) Verb [SC NP1 Prep NP2 Adv ADJ] … (Scontras et al. 2018)

As can be observed in (1), each small clause includes a noun (NP1) that is modified by a prepositional phrase containing a local noun (NP2). The small-clause subject (NP1) agrees with the predicative adjective or participle (ADJ), with the NP2 in between, as a distractor. The sentence is thus only grammatical if agreement of both number and gender takes place between the NP1 and the ADJ.

The stimuli were created by manipulating the number and gender of NP1, NP2, and ADJ, which resulted in grammatical and ungrammatical sentences. Of the 8 stimuli sentences, 4 targeted gender (mis)matches and 4 targeted number (mis)matches. Regarding the gender sentences, half were designed so that the ADJ agreed in both gender and number with the NP1 (grammatical), while matching NP2 only in number. In the other half, the ADJ agreed in both gender and number with the NP2 (ungrammatical), while matching NP1 only in number. In all cases, the two nouns differed in gender but shared the same number, as illustrated in (2).

(2) a. El nen considera l' article de la revista
the kid consider.PRS.3SG the.M/F.SG article.M.SG of the.F.SG magazine.F.SG
completament tràgic.
completely tragic.M.SG

b. *El nen considera l' article de la revista
the kid consider.PRS.3SG the.M/F.SG article.M.SG of the.F.SG magazine.F.SG
completament tràgica.
completely tragic.F.SG
Intended: 'The boy considers the article in the magazine completely tragic.'

The same was done with the sentences targeting number. Two were grammatical, with the ADJ agreeing with NP1 in gender and number, and only in gender with NP2. The other two were ungrammatical, with the ADJ agreeing with the NP2 in gender and number, and only in gender with the NP1. Again, in all cases, the two nouns differed in number but shared the same gender, as in (3).

(3) a. Considero la campana de les escoles excessivament sorollosa.
consider.PRS.1SG the.F.SG bell.F.SG of the.F.PL school.F.PL excessively noisy.F.SG

b. *Considero la campana de les escoles excessivament sorolloses.
consider.PRS.1SG the.F.SG bell.F.SG of the.F.PL school.F.PL excessively noisy.F.PL

Intended: 'I consider the school bell excessively noisy.'

This design allowed us to test whether heritage speakers made more errors in number or in gender, whether these errors were driven by agreement attraction, and whether this pattern was consistent across grammatical and ungrammatical sentences.

3.2.3. Task

The elicited imitation task was administered via a PowerPoint presentation. In each trial, participants first heard a sentence, followed by a short pause during which numbers, figures and letters were displayed on the screen. At the end of this pause, a beep indicated that they could repeat the sentence. This design was chosen because the rationale of this task is that sentence recall does not produce a passive copy but measures implicit linguistic knowledge (Bowles 2011, Ellis 2005, Erlam 2006). Introducing the pause has further been shown to reduce speakers' reliance on working memory (Kostromitina and Plonsky 2022).

The task included two practice sentences (one grammatical, one ungrammatical) to familiarize participants with the procedure. The stimulus sentences were elicited by the experimenter, a mainland Catalan speaker, and were presented in a randomized order to prevent confounding learning effects. Sentences were repeated once, and only twice if requested.

3.2.4. Procedure

The task was administered online via Google Meet. Participants were instructed to use a computer in a quiet environment with a stable internet connection. Sessions were audio-recorded using OBS Studio, and later transcribed and coded for analysis. At the end of each session, demographic information and linguistic history were collected from participants using the survey platform Qualtrics (https://www.qualtrics.com). The demographic information included participants' age, gender, place of birth, and place of residence. The linguistic information was divided into two sections: the first section consisted of general information about language acquisition, fluency, and use, while the second section consisted of information about Catalan and Dutch or German, which included schooling and proficiency levels.

Participants took part in the experiment on a voluntary basis and gave oral consent at the beginning of each session. The study received positive post-hoc advice from the Linguistics Chamber of the Faculty Ethics assessment Committee of the Faculty of Humanities (FEtC-H) (reference number: 25-076-02).

At the beginning of the session, participants were provided with a detailed explanation of the task procedure and were explicitly informed that some sentences might sound strange or unusual, and that they should repeat them as naturally as possible. During the practice phrase, when participants failed to correct the ungrammatical sentence, they were asked if they had noticed anything strange and how they would normally say it. When participants did not correct it, the experimenter simply confirmed that this was the intended approach. This was done following observations from the pilot study, in which participants tended to repeat ungrammatical sentences without correcting them, even when they were aware of the errors, as they interpreted the instructions as requiring verbatim repetition. After the two practice trials, we addressed any remaining questions before starting the main task.

	Mainland speakers (n= 8)			First-generation speakers (n = 19)			Heritage speakers (n = 9)		
	Count	*M* (%)	*SD*	Count	*M* (%)	*SD*	Count	*M* (%)	*SD*
Repeat grammatical	32/32	100.0	0.0	73/75	96.9	9.3	26/33	72.2	42.3
Correct ungrammatical	30/32	93.8	11.6	69/76	90.8	17.1	24/29	82.8	22.7

Table 2: Performance in repeating grammatical sentences and correcting ungrammatical sentences.

During the main task, to mitigate the impact of unstable internet connection, the experimenter always made sure that all stimuli were heard clearly. Participants were instructed to ask for a repetition if any part of a stimulus was unclear or inaudible. Despite this, one participant was unable to complete the task due to auditory difficulties.

3.3. Results

This section presents the results of the elicited imitation task. We begin with an overview of speakers' overall performance, followed by a general overview of the error types, and end with an analysis of agreement attraction effects based on the features of the head noun. Given the limited number of stimuli, we will report the raw counts without statistical results.

3.3.1. Overall performance

The performance of the three participant groups in repeating the grammatical and ungrammatical stimuli is summarized in table 2. For grammatical sentences, responses were considered correct if the resulting sentence was grammatical, whether the participants produced an exact repetition or modified the number and/or gender features of NP1, NP2 and/or ADJ while remaining grammatical. For ungrammatical sentences, responses were considered correct only when participants modified the sentence to make it grammatical. Instead, exact repetitions or changes that lead to an ungrammatical sentence were counted as incorrect. Note that only those sentences that maintained the structure in (1) were kept in the analysis. Changes that led to a different syntactic construction were discarded.

Overall, table 2 shows that mainland speakers, independently of sentence type, achieved the highest accuracy. First-generation speakers were less accurate in their responses, but with comparable results to mainland speakers. In contrast, heritage speakers showed a somewhat lower accuracy, especially when repeating grammatical sentences, where their performance varied considerably.

The relatively large standard deviations observed in some conditions, particularly among heritage speakers, indicate substantial individual variation in overall task performance. Such variability is extensively reported in heritage language populations and is often linked to differences in language exposure, use, and proficiency, among others, which can lead to individual variation exceeding between-group variation (e.g., Özsoy and Blum 2023). Importantly, however, this variation is related to overall accuracy rather than the structural patterns examined in the present study, as it will be shown that the effects related to agreement attraction remain consistent across participants.

To better understand participants' performance, table 3 presents the nature of their responses for each sentence type. Specifically, it shows whether they produced exact repetitions or changed them, and whether those changes resulted in grammatical or ungrammatical outputs. Note that while exactly repeating a grammatical sentence yields a grammatical output, exactly repeating an ungrammatical sentence does not.

Table 3 restates what is observed in table 2, but in greater detail. Mainland speakers show near-ceiling performance: they produced a high percentage of exact repetitions of grammatical sentences and corrected almost all ungrammatical sentences. First-generation speakers followed a similar pattern: they were very accurate in their repetitions of grammatical sentences and corrected most errors. Their only frequent error (although at a very low percentage) was the exact repetition of ungrammatical sentences. In contrast, heritage speakers showed more variability in their data: they generally produced grammatical outputs but struggled more with exact repetition than the other groups. When faced with ungrammatical sentences, they

	Mainland speakers			First-gen speakers			Heritage speakers		
	Count	*M* (%)	*SD*	Count	*M* (%)	*SD*	Count	*M* (%)	*SD*
Grammatical sentence									
Exact repetition	24/32	75.0	18.9	58/75	76.3	27.0	15/33	41.7	41.5
Changed grammatical	8/32	25.0	18.9	15/75	20.6	21.4	11/33	30.6	34.9
Changed ungrammatical	0/32	0.0	0.0	2/75	3.1	9.3	7/33	27.8	42.3
Ungrammatical sentence									
Exact repetition	2/32	6.3	11.6	6/76	7.9	14.6	2/29	9.4	18.6
Changed grammatical	30/32	93.6	11.6	69/76	90.8	17.1	24/29	80.2	22.7
Changed ungrammatical	0/32	0.0	0.0	1/76	1.3	5.7	3/29	10.4	14.6

Table 3: Response types for grammatical and ungrammatical sentences.

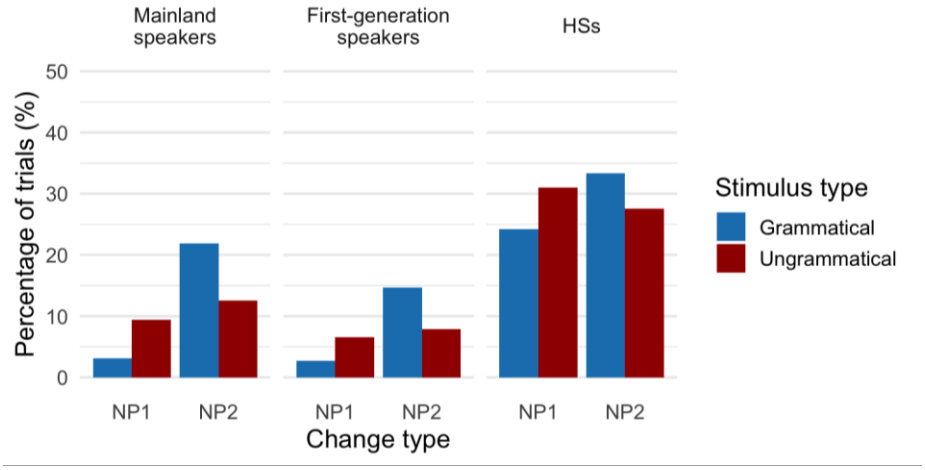

Figure 1: NP1 and NP2 changes grouped by stimulus type.

generally corrected them, but they were just as likely to repeat the error as they were to change it for a different ungrammatical version.

Finally, an important aspect to consider for the analysis is whether participants only changed the number and gender features of the adjective, as expected, or whether they also modified the features of NP1 and/or NP2 in their production. Figure 1 illustrates these changes. Heritage speakers produced the highest proportion of changes overall, followed by mainland speakers, with first-generation speakers producing the lowest proportion. Interestingly, across groups, NP2 was more often changed in grammatical than ungrammatical stimuli, while NP1 showed the opposite pattern, being more often changed in ungrammatical than grammatical stimuli.

3.3.2. Agreement errors type

In this section, we examine the agreement errors in gender and number. These comprise agreement attraction errors, which involve these phi-features, as well as gender errors that follow a masculine-default strategy or a shape-based strategy. The masculine default strategy consists of assigning masculine gender to all nouns, regardless of whether they are masculine or feminine. The shape-based strategy consists of assigning gender on the basis of the phonological, morphological, or orthographic form of the noun (e.g., in Spanish, nouns ending in *-o* are associated with masculine and those ending in *-a* with feminine). The distribution of gender and number error types is summarized in table 4.

	Mainland speakers		First-generation speakers		Heritage speakers	
	Count	%	Count	%	Count	%
Gender errors						
Agreement attraction	1/64	1.6	5/151	3.3	3/62	4.8
Masculine default	0/64	0.0	1/151	0.7	2/62	3.3
Shape-based	0/64	0.0	0/151	0.0	4/62	6.5
Number errors						
Agreement attraction	1/64	1.6	3/151	2.0	1/62	1.6

Table 4: Gender and number error types (counts and proportions).

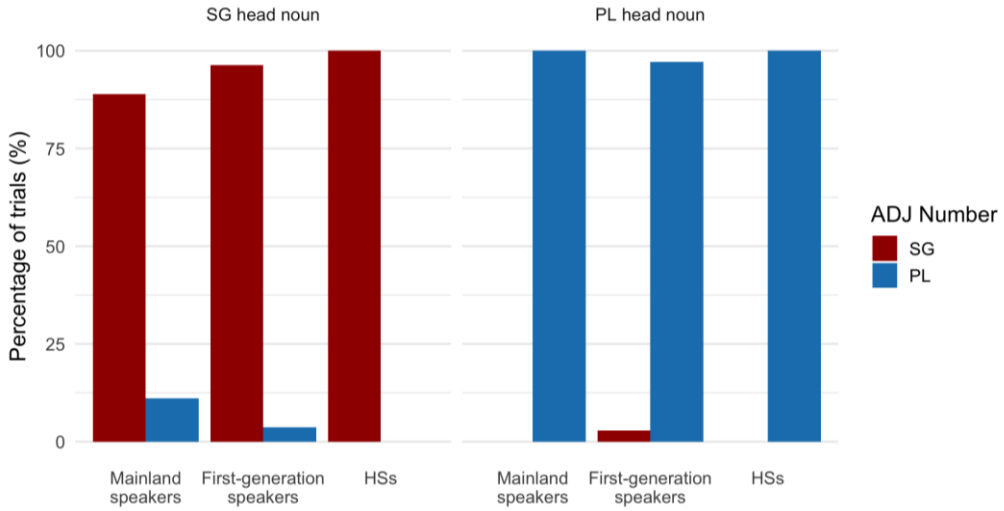

Figure 2: Speakers' production of singular and plural adjectives grouped by NP1 number.

Overall, gender and number errors were limited across speaker groups, with heritage speakers producing more errors, although at low rates. Across groups, gender-related errors were more frequent than number ones. Among the error subtypes, agreement attraction was the most common: it was the only error type produced by mainland speakers and the most frequent one among first-generation speakers. For heritage speakers, however, shape-based errors showed the highest percentage, despite the differences in raw counts being minimal.

3.3.3. Agreement attraction

This section follows Fuchs et al. (2015) and Scontras et al. (2018) in examining agreement attraction effects for number and gender, with the analysis organized by the features of the head noun (NP1), which determines agreement. In order to compare the results of these studies to those of the present one, we will only use those sentences where participants only changed the features of the ADJ in their repetitions, and not those of the NP1 and/or NP2. Based on this criterion, figure 2 shows participants' production of adjective number as a function of NP1 number, while figure 3 presents their production of adjective gender as a function of NP1 gender.

Singular head noun: To avoid possible effects of gender mismatches, gender was kept constant across NP1, NP2, and ADJ. To elicit agreement attraction, NP2 was assigned plural number, in contrast to the singular NP1. Thus, all cases in which the ADJ appeared in the plural reflect agreement attraction. As shown in Figure 1, all groups produced a high percentage of singular adjectives, consistent with the head noun. Performance was at ceiling in the case of heritage speakers. Both mainland and first-generation

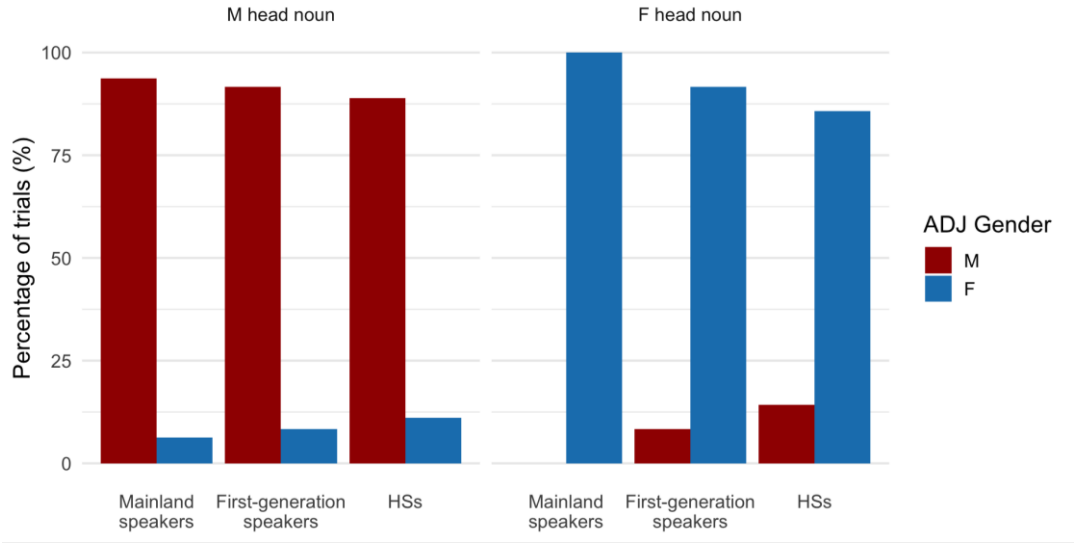

Figure 3: Speakers' production of masculine and feminine adjectives grouped by NP1 gender.

speakers produced plural adjectives due to agreement attraction effects, but the percentage was low. Despite the difference being minimal, mainland speakers were slightly outperformed by the other two groups in this condition.

Plural head noun: As in the singular head noun condition, gender was fixed across NP1, NP2, and ADJ. In this case, the NP2 was singular, in opposition to the head noun. All groups produced a very high percentage of plural adjectives, in agreement with the NP1. Performance was at ceiling for mainland and heritage speakers, with only a few number errors produced by first-generation speakers.
Masculine head noun: To avoid possible effects of number mismatches, number was kept constant across NP1, NP2 and ADJ. To trigger agreement attraction, NP2 was assigned feminine gender, in contrast to the masculine NP1. Consequently, all the cases where the ADJ appeared in the feminine reflect agreement attraction. As can be observed in figure 2, all three groups produced a high percentage of masculine adjectives, with only a small proportion of feminine errors.

Feminine head noun: As with the analysis of masculine head nouns, number was fixed across NP1, NP2 and ADJ. In this case, NP2 was masculine, in contrast with the feminine head noun. Across speaker groups, the percentage of correct agreement between the NP1 and the ADJ was high, reaching ceiling levels for mainland speakers. Both heritage speakers and, to a lesser extent, first-generation speakers produced agreement attraction errors by using a masculine ADJ, although the percentages are low.

4. Discussion, shortcomings, and conclusions

This study departs from the widely recognized challenge that the design and selection of experimental tasks crucially shape the data elicited from speakers of non-standardized varieties. Acceptability judgment tasks, although widely used to probe grammatical knowledge, have been shown to be sensitive to task parameters in ways that can significantly affect outcomes (Marty et al. 2020). To address this issue, we compare two methodologies—acceptability judgments and elicited imitation—to explore heritage speakers' grammatical representations.

Our starting point is the study by Scontras et al. (2018), who investigated gender and number agreement in Spanish heritage speakers in the US using an acceptability judgment task. Their results suggested differences between heritage speakers and monolinguals: while monolinguals rated sentences with a single feature mismatch (gender or number) as more acceptable than those with mismatches in both features, heritage speakers rated both types of violations equally. This was interpreted as evidence that, in

heritage grammars, gender and number features are bundled and valued together. Moreover, they found asymmetries in feature content: both monolinguals and heritage speakers tolerated default masculine adjectives with feminine nouns, but only heritage speakers accepted singular adjectives with plural nouns. This was taken as evidence of feature loss in number, with only plural specified in heritage grammars.

To test whether such findings may be task-dependent, we conducted an elicited imitation task, which is argued to tap into implicit grammatical knowledge. Following the stimulus structure of Scontras et al. (2018) while focusing specifically on feature content, we manipulated gender and number separately across the head noun (NP1), intervener (NP2), and adjective (ADJ), controlling for potential confounds. Contrary to previous findings, our results revealed no evidence of feature loss: heritage Catalan speakers made very few errors, at rates comparable to first-generation and mainland speakers, and they frequently corrected ungrammatical sentences regardless of the features involved. These findings indicate that heritage grammars retain separate values for both number and gender.

The discrepancy between our findings and generalizations derived from previous judgment-based studies suggests that the "incomplete" label often applied to heritage contexts may be an artifact of methodology. While heritage speakers may appear inconsistent in explicit ratings, their performance in the repetition task reveals a systematicity in their morphosyntactic representations. As seen in other heritage contexts (e.g., Tokaç-Scheffer et al. 2023, Van Baal 2025), a speaker may fail to explicitly reject an ungrammatical form while still demonstrating a robust, real-time sensitivity to those same grammatical features during production. As Bayram et al. (2019) point out, a grammar should not be labeled "incomplete" simply because it differs from an arbitrary standard variety. Our study shows that when using an implicit task (repetition), heritage speakers show high accuracy in number and gender agreement, confirming that these features are indeed retained and systematically represented in their minds, even if their performance differs from mainland speakers in other experimental settings. These results support a shift away from deficit-based narratives toward a view of heritage grammars as dynamic and internally consistent systems.

This study has a number of limitations that were listed throughout. First, given the limited literature on Catalan as a heritage language, we considered Spanish as a point of departure, as the syntax of the two languages is similar as far as gender is concerned. Then, the number of speakers is very small, also due to the fact that the Catalan community is of recent immigration. This means that the generalizations are only tentative and cannot be considered 100% accurate.

All in all, however, these outcomes highlight two key points. First, they challenge deficit-based accounts by showing that heritage speakers' feature representations can be robust when tested with tasks targeting implicit knowledge. Second, they demonstrate that methodological choices strongly affect the patterns we observe: while judgment tasks may underestimate heritage competence, repetition-based tasks can reveal richer aspects of grammatical representation. We therefore argue for a multi-method approach in heritage language research, where complementary methodologies are combined to provide a more comprehensive picture of heritage grammars. Only through such an approach can robust theoretical generalizations be drawn and simplistic deficit models be overcome.

Acknowledgments

To Johan: an inspiration, a champion of freedom, a friend, with admiration and affection.

Parafita Couto acknowledges support from The Babel Brain: Mapping Multilingual Ecologies (BabelBrain), ATR2024-154931 funded by MICIU/AEI/10.13039/501100011033.

References

Albirini, Abdulkafi, Elabbas Benmamoun, and Brahim Chakrani. 2013. Gender and number agreement in the oral production of Arabic heritage speakers. *Bilingualism: Language and Cognition* 16 1: 1–18. https://doi.org/10.1017/S1366728912000132.

Andriani, Luigi, Jan Casalicchio, Francesco Maria Ciconte, Roberta D'Alessandro, Alberto Frasson, Brechje van Osch, Luana Sorgini, and Silvia Terenghi. 2022. Documenting Italo-Romance minority languages in the Americas. Problems and tentative solutions. In *Contemporary Research in Minority and Diaspora Languages of Europe*, edited by Andrew Nevins and Matt Coler, no. 6 in Contact and Multilingualism, pp. 9–56. Language Science Press, Berlin. https://doi.org/10.5281/zenodo.4902965.

van Baal, Yvonne. 2025. Acceptability judgments in moribund heritage languages: Mitigating the challenges. *Bergen Language and Linguistics Studies* 15 1: 7–14. https://doi.org/10.15845/bells.v15i1.4548.

Bayram, Fatih, Tanja Kupisch, Diego Pascual y Cabo, and Jason Rothman. 2019. Terminology matters on theoretical grounds too! Coherent grammars cannot be incomplete. *Studies in Second Language Acquisition* 41 2: 257–64. https://doi.org/10.1017/S0272263119000287.

Busterud, Guro, Terje Lohndal, Toril Opsahl, Yulia Rodina, and Marit Westergaard. 2025. Language change and the loss of feminine gender: Grammatical gender and declension class in the Oslo dialect. *Nordic Journal of Linguistics*: 1–26. https://doi.org/10.1017/S0332586525000071.

Colina Fortuny, Laia. 2025. *Object Pronouns in Catalan as a Heritage Language in Germany and in the Netherlands*. Master's thesis, Utrecht University. Available at https://studenttheses.uu.nl/handle/20.500.12932/50261.

Ellis, Rod. 1991. Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition* 13: 161–86. https://doi.org/10.1017/S0272263100009931.

D'Alessandro, Roberta, David Natvig, and Michael T. Putnam. 2021. Addressing challenges in formal research on moribund heritage languages: A path forward. *Frontiers in Psychology* 12: 700126. https://doi.org/10.3389/fpsyg.2021.700126.

Fuchs, Zuzanna, Maria Polinsky, and Gregory Scontras. 2015. The differential representation of number and gender in Spanish. *The Linguistic Review* 32 4: 703–37. https://doi.org/10.1515/tlr-2015-0008.

Jegerski, Jill and Sara Fernández Cuenca. 2025. Variability in the online processing of subject-verb number agreement in Spanish as a heritage language: The role of lexical frequency. *Languages* 10 9: 211. https://doi.org/10.3390/languages10090211.

Leivada, Evelina, Roberta D'Alessandro, and Kleanthes Grohmann. 2019. Eliciting big data from small, young, or non-standard languages: 10 Experimental challenges. *Frontiers in Psychology* 10: 313. https://doi.org/10.3389/fpsyg.2019.00313.

Leivada, Evelina, Itxaso Rodríguez-Ordóñez, M. Carmen Parafita Couto, and Silvia Perpiñán. 2023. Bilingualism with minority languages: Why searching for unicorn language users does not move us forward. *Applied Psycholinguistics* 44 3: 384–99. https://doi.org/10.1017/S0142716423000036.

Lohndal, Terje and Marit Westergaard. 2016. Grammatical gender in American Norwegian Heritage Language: Stability or attrition? *Frontiers in Psychology* 7: 344. https://doi.org/10.3389/fpsyg.2016.00344.

Lohndal, Terje and Marit Westergaard. 2021. Grammatical gender: Acquisition, attrition, and change. *Journal of Germanic Linguistics* 33 1: 95–121. https://doi.org/10.1017/S1470542720000057.

Marty, Paul, Emmanuel Chemla, and Jon Sprouse. 2020. The effect of three basic task features on the sensitivity of acceptability judgment tasks. *Glossa: a journal of general linguistics* 5 1: 72. https://doi.org/10.5334/gjgl.980.

Montrul, Silvina. 2002. Incomplete acquisition and attrition of Spanish tense/aspect distinctions in adult bilinguals. *Bilingualism: Language and Cognition* 5 1: 39–68. https://doi.org/10.1017/S1366728902000135.

Montrul, Silvina. 2008. *Incomplete Acquisition in Bilingualism: Re-examining the Age Factor*. No. 39 in Studies in Bilingualism. John Benjamins, Amsterdam. https://doi.org/10.1075/sibil.39.

Özsoy, Onur and Frederic Blum. 2023. Exploring individual variation in Turkish heritage speakers' complex linguistic productions: Evidence from discourse markers. *Applied Psycholinguistics* 44 4: 534–64. https://doi.org/10.1017/S0142716423000267.

Özsoy, Onur, Kateryna Iefremenko, and Christoph Schroeder. 2022. Shifting and expanding clause combining strategies in heritage Turkish varieties. *Languages* 7 3: 242. https://doi.org/10.3390/languages7030242.

Polinsky, Maria. 2006. Incomplete acquisition: American Russian. *Journal of Slavic Linguistics* 14: 161–219.

Polinsky, Maria. 2008. Gender under incomplete acquisition: Heritage speakers' knowledge of noun categorization. *Heritage Language Journal* 6 1: 40–71.

Polinsky, Maria. 2018. *Heritage Languages and Their Speakers*. No. 159 in Cambridge Studies in Linguistics. Cambridge University Press, Cambridge. https://doi.org/10.1017/9781107252349.

Putnam, Michael T. and Liliana Sánchez. 2013. What's so incomplete about incomplete acquisition? A prolegomenon to modeling heritage language grammars. *Linguistic Approaches to Bilingualism* 3 4: 478–508. https://doi.org/10.1075/lab.3.4.04put.

Rodríguez, Estrella and Lara Reglero. 2015. Heritage and L2 processing of person and number features: Evidence from Spanish subject-verb agreement. *EuroAmerican Journal of Applied Linguistics and Languages* 2 2: 11–30. https://doi.org/10.21283/2376905X.3.46.

Romano, Francesco and Pedro Guijarro-Fuentes. 2023. Task effects and the yes-bias in heritage language bilingualism. *International Journal of Bilingual Education and Bilingualism* 27 3: 389–409. https://doi.org/10.1080/13670050.2023.2206949.

Rothman, Jason, Fatih Bayram, Vincent DeLuca, Grazia Di Pisa, Jon Andoni Duñabeitia, Khadij Gharibi, Jiuzhou Hao, Nadine Kolb, Maki Kubota, Tanja Kupisch, Tim Laméris, Alicia Luque, Brechje van Osch, Sergio Pereira Soares, Yanina Prystauka, Deniz Tat, Aleksandra Tomić, Toms Voits, and Stefanie Wulff. 2023. Monolingual comparative normativity in bilingualism research is out of "control": Arguments and alternatives. *Applied Psycholinguistics* 44 3: 316–29. https://doi.org/10.1017/S0142716422000315.

Scontras, Gregory, Zuzanna Fuchs, and Maria Polinsky. 2018. In support of representational economy: Agreement in heritage Spanish. *Glossa: a journal of general linguistics* 3 1: 1. https://doi.org/10.5334/gjgl.164.

Sprouse, Jon and Diogo Almeida. 2017. Setting the empirical record straight: Acceptability judgments appear to be reliable, robust, and replicable. *Behavioral and Brain Sciences* 40: e311. https://doi.org/10.1017/S0140525X17000590.

Sprouse, Jon, Carson Schütze, and Diogo Almeida. 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. *Lingua* 134: 219–48. https://doi.org/10.1016/j.lingua.2013.07.002.

Tokaç-Scheffer, Suzan D., Seçkin Arslan, and Lyndsey Nickels. 2023. Insights into the time course of evidentiality processing in Turkish heritage speakers using a self-paced reading task. *Frontiers in Communication* 8: 1070510. https://doi.org/10.3389/fcomm.2023.1070510.