Available online at www.sciencedirect.com

ScienceDirect

Lingua 318 (2025) 103911

Lingua

www.elsevier.com/locate/lingua

Bilingual influences and sources of variability in acceptability judgments: A case study of Chinese

Hai Hu [a,*,1], Aini Li [b,1], Yina Patterson [c], Jiahui Huang [d], Chien-Jer Charles Lin [e]

[a] School of Foreign Languages, Shanghai Jiao Tong University, China
[b] Department of Linguistics and Translation, City University of Hong Kong, China
[c] Department of Asian and Near Eastern Languages, Brigham Young University, USA
[d] Department of Linguistics, University of Washington, USA
[e] Department of East Asian Languages and Cultures, Indiana University Bloomington, USA

Received 31 July 2024; revised 10 February 2025; accepted in revised form 11 February 2025;

Abstract

The replicability of grammaticality judgments forms the foundation of data quality in linguistic research. Previous work has mostly focused on judgments from ideal "native speakers," disregarding speakers of different language backgrounds. This study examines whether acceptability judgments in Chinese can be replicated by linguistically diverse native speakers, "monodialectal" and "multidialectal" speakers of Chinese, and then explores how various factors influence such judgments. First, we obtained a representative dataset by randomly sampling 337 minimal pairs from 68 journal articles on Chinese syntax from the past decade. Then, two groups of participants—monolingual Mandarin speakers from Beijing and Mandarin-Cantonese bilinguals from Guangzhou—completed an acceptability rating task (Experiment 1). Two forced-choice experiments (Experiments 2 and 3) were conducted to further examine the unreplicated pairs from Experiment 1. The results of these three experiments showed a convergence rate of 92% between our participants and the syntacticians who authored the examples. Importantly, the language backgrounds of the participants and the authoring syntacticians were not found to play a role in acceptability judgments, whereas sentence length and the language of the journals did. The multilingual status of Cantonese-Mandarin bilinguals has a subtle but limited influence on judgments in Mandarin Chinese. We argue that the reliance on a monolingual "ideal" native speaker for eliciting judgments may have been overemphasized in linguistic research.

Keywords: Acceptability judgments; Experimental syntax; Bilingualism; Variability; Chinese

* Corresponding author.
E-mail addresses: hu.hai@sjtu.edu.cn (H. Hu), ainili@cityu.edu.hk (A. Li), yina_patterson@byu.edu (Y. Patterson), huangjh@uw.edu (J. Huang), chiclin@iu.edu (C.-J.C Lin).
[1] Co-first authors with equal contributions.

https://doi.org/10.1016/j.lingua.2025.103911

1. INTRODUCTION

Recognizing the dichotomy between linguistic competence and performance is fundamental to modern linguistic studies. The intuition of native speakers on whether sentences are grammatical or not forms the basis of syntactic theories of linguistic competence. In general, in theoretical studies, these acceptability judgments are informal in nature, as they are obtained through introspection rather than through carefully controlled experiments. In contrast to such informal judgments, a more rigorous experimental procedure should ideally aggregate the judgments made by many native speakers on stimuli in a controlled setting.

In recent years, many studies have been conducted to verify whether informal judgments described in the syntactic literature can be replicated by large groups of speakers, usually linguistically naïve native speakers of English (Sprouse et al., 2012; Sprouse et al., 2013; Mahowald et al., 2016; Myers, 2009). There have also been studies on other languages; for instance, Weskott and Fanselow, 2011 focusing on German, Linzen and Oseki, 2018 examining Japanese and Hebrew, and Chen et al., 2020 focusing on Chinese, among others. However, studies on non-English languages have been relatively limited, and more comprehensive studies are called for. To the best of our knowledge, this study is the first to examine the acceptability judgments of non-English sentences using a broad and representative sample from a wide range of topics, authors, and publication venues.

While the ultimate goal of syntactic research is to isolate the exclusively grammatical factors underlying acceptability judgments, an array of factors not directly related to grammatical competence may come into play when making judgments about the well-formedness of sentences. Schütze (2016) classified these factors into task-related factors (e.g., context, parsability) and subject-related factors (e.g., dialectal backgrounds of the participants/authors, participants' education level, age, gender). In terms of task-related factors, previous studies have shown that processing effects, such as repeated exposure effects (Chaves and Dery, 2014; Chaves and Jeruen, 2018; Francom, 2009; Hofmeister and Norcliffe, 2013; Lin, 2018; Snyder, 2000), and individual differences in working memory capacity (Hofmeister et al., 2012; Phillips, 2013; Sprouse et al., 2012) can influence acceptability judgments. Moreover, complex sentences often present parsing difficulties for participants and are more likely to be judged as ungrammatical (Bever, 1970; Yao et al., 2022). In addition to processing factors, semantic and pragmatic factors also affect syntactic judgments. Regarding Chinese, for instance, it has been argued that the naturalness of a sentence is, to a significant degree, dependent on factors such as the discourse and presuppositions triggered by pragmatic markers of a sentence (Yao et al., 2022).

Of particular interest in the current study are subject-related factors, specifically sociolinguistic factors, and the role of language background, which have often been overlooked in previous research. In fact, we do not see systematic discussions in current theoretical studies of variation in the judgments of speakers with different dialects and backgrounds, although it has been discussed as a potential factor in previous literature (Chen et al., 2020; Linzen and Oseki, 2018); it has also been observed that acceptability judgments may differ for speakers of different varieties/dialects of a language. For instance, the sentence "The car needs washed" is considered acceptable by some speakers of American English but not by others (e.g., Murray et al., 1996; Edelstein, 2014). Nevertheless, apart from a few exceptions (Zanuttini et al., 2018; Barbiers and Bennis, 2007; Poletto and Benincà, 2007), very few empirical studies have quantified this variation. Crucially, we need to estimate the magnitude of such variations, as they will have theoretical and practical consequences in (experimental) syntax research for recruiting subjects. That is, should we only trust an ideal native speaker (e.g., a monolingual speaker of Beijing Mandarin in the case of Mandarin Chinese)? Is it equally acceptable to elicit judgments from speakers from multilingual backgrounds? How do we define "native intuition"? Related to speaker background, a comment we often hear in syntax classrooms or about syntactic research is how the backgrounds of researchers may also influence the reliability of acceptability judgments. In the Chinese context, this may refer to authors using different varieties of Mandarin Chinese (Northern Mandarin, Southern Mandarin, Taiwan Mandarin (Guoyu), etc.).

To better understand how language/dialectal differences influence acceptability judgments, we recruited two groups of participants who are speakers of Mandarin Chinese (living in Beijing and Guangzhou) with vastly different language/dialectal backgrounds. Participants from Beijing speak Beijing Mandarin as their first language, whereas those from Guangzhou are balanced bilingual speakers of Cantonese and Standard Mandarin.

Within the Sino-Tibetan language family, seven languages are commonly distinguished in the Sinitic branch: Mandarin, Min, Yue, Wu, Hakka, Gan, and Xiang (Li, 1973). Chinese linguists generally agree that there exists a primary split between northern Sinitic languages such as Mandarin and southern languages (to the south of the Yangtze River). While Mandarin (i.e., Beijing Mandarin) belongs to the northern Chinese branch, Yue (i.e., Cantonese) belongs to the southern branch. Linguistic varieties within the Mandarin branch are usually mutually intelligible, whereas those in the southern branch are not (Tang and Van Heuven, 2009). Such differences have been claimed to be primarily phonological or lexical. Tang and van Heuven (2007) and Tang and Van Heuven (2009) used experimental means (both opinion and functional tests) to test the extent to which Chinese varieties are mutually intelligible. According to these

Table 1
Summary of the three experiments, each completed by both Beijing and Guangzhou participants.

Experiment No.	Task	Items included
1	Acceptability rating task on 7-point Liker scale	337 pairs (in 6 lists) and 2 catch trials
2	Forced-choice task	34 test pairs, 17 control pairs, 2 catch trials
3	Forced-choice task	26 test pairs, 17 control pairs, 2 catch trials

studies, Beijing listeners could correctly recognize only 34% of the words spoken by speakers from Guangzhou. Therefore, based on the degree of mutual intelligibility with Mandarin, Cantonese is treated as a language rather than a dialect in the current study.[2]

The goal of this study is twofold. *First*, by retrieving 337 minimal pairs from 68 journal articles on Chinese syntax (broadly defined) from 10 journals published in Chinese or English, written by authors from different Chinese-speaking communities—the Chinese mainland, Hong Kong, Taiwan, and elsewhere—we aim to provide a large sample of judgments on Mandarin Chinese from sources representing active syntactic research. Using this dataset, we examine the reliability of the authors' judgments by conducting an acceptability judgment experiment, based on a 7-point Likert scale (LS), with a large pool of Mandarin Chinese speakers. We then further examine problematic sentence pairs from the LS experiment using a forced-choice (FC) task. *Second*, to quantify the extent to which the dialectal/language background of the participants may influence their judgments, we recruited participants from two regions with distinctive language backgrounds: Beijing and Guangzhou (Canton). The Beijing participants speak Beijing Mandarin, serving as idealized monolingual native speakers in the Chinese context. The Guangzhou participants are all bilingual speakers, fluent in both Cantonese and Mandarin Chinese; their judgments will be a window for us to examine how Cantonese influences their judgments on Mandarin Chinese and crucially, how different their judgments are from those of the Beijing participants. We also examine whether the example sentences created by syntacticians from various Chinese-speaking communities lead to different judgments.

In summary, we consider the following research questions:

1. How do the acceptability judgments made by Cantonese-Mandarin bilinguals differ from those made by (monolingual) Beijing Mandarin speakers?
2. How reliable are the judgments made for Chinese sentences by authors of journal articles compared with judgments obtained from native language users in an experimental setting?
3. What factors may play a role in acceptability judgments, including, for example, sentence length, article language, and authors' regions of origin?

To address these research questions, we conducted three experiments. In **Experiment 1**, we asked participants to rate the naturalness of sentences (on a 7-point LS) to examine whether acceptability judgments from journal articles could be replicated and whether monolingual and bilingual participants differ in their judgments. Sentences that were not replicated in Experiment 1 were further tested in **Experiments 2 and 3** using FC tasks to examine, in a more direct way, how participants would rate each sentence against its minimally different counterpart in a given contrast, and whether monolingual and bilingual participants differ in this respect. An overview of the experiments is presented in Table 1.

The remainder of the paper is structured as follows. In Sections 2–4, we detail the process of stimuli curation, participant recruitment, experimental procedures, and the results of Experiments 1, 2, and 3, respectively. Section 5 discusses the results, the replicability of judgments in general, the influence of language background on the judgments, and the implications of our results. Finally, Section 6 concludes the paper.

Note that the goal of our study is to assess the overall reliability of linguists' informal grammaticality judgments in published articles rather than examine the soundness of a specific syntactic theory or analysis. Thus, we will refrain from singling out specific syntactic phenomena for Chinese and only mention a few examples in Section 5.4. We have made our data and analysis scripts available at https://osf.io/z5pts/ and included all minimal pairs that our experiments did not replicate in the repository. Readers interested in specific pairs are encouraged to check the judgments presented there.

[2] In this paper, we refer to the Beijing participants as "monolingual" Mandarin speakers, and the Guangzhou participants as Cantonese-Mandarin "bilingual" speakers. We understand that they may not be strictly "monolingual" or "bilingual," as many participants are students in China, who are required to study English from primary school to college. However, they are mostly monolingual or bilingual for Chinese languages. We collect detailed information on their language background, reported in Section 2.2.

2. EXPERIMENT 1

2.1. Stimuli curation

We sampled articles from 10 academic journals in the field of linguistics that publish research articles on Chinese syntax. The journals were selected to be diverse in author background, journal languages (i.e., English versus Mandarin Chinese), publishers, etc., as shown in Table 2. A total of 68 articles were selected. We sampled more articles from the two journals that publish in Chinese (*yǔyán kēxué* (*Language Sciences*) and *zhōngguó yǔwén* (*Chinese Philology*)) in order to balance the numbers of articles that were published in Chinese and English.

Next, we copied all example sentences from the 68 articles into a spreadsheet, including those in the footnotes, resulting in approximately 7,000 sentences in total. For each paper, we randomly sampled six example sentences that were deemed ungrammatical by the author(s); if a paper had fewer than six such sentences, we sampled all of them. "Ungrammatical" is operationalized as any example sentence marked with ∗, ∗?,?∗, or ??; thus, those marked with a single? are excluded. This process resulted in 397 ungrammatical sentences in total. Similar to what has been done in previous literature, for example, Sprouse et al. (2013), we then excluded those that are marked ungrammatical for any of the following reasons as they would be difficult to assess in our planned text-based acceptability judgment task: (a) prosody, (b) information statuses involving focus/topic, (c) anaphoric relations, (d) *pro*, (e) whether a specific reading/interpretation of the sentence is unavailable. Examples illustrating the first four categories are provided below.

1 Example removed due to specific intended focus. This is removed because it is difficult to indicate focus in a text-based experimental setting.

1a	Text	zhāng sān	shuǐguǒ	**shénme**	zuì	cháng	chī?
	Gloss	John	fruit	WHAT	most	often	eat
	Trnsltn.	'What is the fruit that John most often eat?' (focus on WHAT)					
1b	Text	*zhāngsān	**shuǐguǒ**	shénme	zuì	cháng	chī?
	Gloss	John	FRUIT	what	most	often	eat
	Trnsltn.	'Fruit, what is it that John most often eat?' (focus on FRUIT)					

2 Example removed due to anaphoric relations or intended reading. This is excluded because it is difficult to know which interpretation the participants are using.

Text	*yuehàn$_i$	bǐ	mǎlì	rènwéi	tā$_i$	gāo.
Gloss	John$_i$	*THAN*	Mary	think	3sg$_i$	tall
Trnsltn.	'John is taller than Mary thinks he is.' (Intended Reading)					

3 Example removed due to *pro*. This is excluded because linguistically naïve speakers do not know what *pro* is.

Text	*lǐsì,	zhāngsān$_i$	kū	dé	*pro*$_i$	hěn	shāngxīn.
Gloss	Lisi,	Zhangsan$_i$	cry	DE	*pro*$_i$	very	sad
Trnsltn.	'Zhangsan cried very sadly for Lisi.'						

After the sampling process, 337 ungrammatical example sentences remained (see Table 2 for information on the sampled sentences and journals). Of these 337 ungrammatical sentences, 322 are marked with *, 11 marked with ??, 2 marked with?*, and 2 marked with ??/*.[3]

[3] A reviewer pointed out that bad sentences with a "?" mark—"??", "?*", or "*?"—may be infelicitous or inappropriate rather than ungrammatical. Here we follow previous literature such as Sprouse et al. (2013) to include these examples, and only exclude the ones marked with a single "?". The bad sentences in this paper are thus mostly considered to be ungrammatical by the linguists who authored the examples, but potentially with a few infelicitous or inappropriate ones.

Table 2
Information on sampled sentences and journals. Note: in the table, “n articles” stands for “the number of articles,” and “n pairs” refers to “the number of pairs.”

Abbr.	Journal	Language	N articles	N pairs
CSL	Concentric: Studies in Linguistics	English	5	19
JCL	Journal of Chinese Linguistics	En + Ch	5 + 1	35
JEAL	Journal of East Asian Linguistics	English	6	26
LI	Linguistic Inquiry	English	5	24
LL	Language and Linguistics	English	6	31
LS	Lingua Sinica	English	6	33
NLLT	Natural Language and Linguistic Theory	English	6	29
TL	Taiwan Journal of Linguistics	English	7	36
yykx	yǔyán kēxué (Language Sciences)	Chinese	10	51
zgyw	zhōngguó yǔwén (Chinese Philology)	Chinese	11	53
Sum			68	337

Our next pre-processing step was to construct minimal pairs for these 337 ungrammatical sentences. Again, following previous literature (Sprouse et al., 2013; Chen et al., 2020), we paired each ungrammatical sentence with a grammatical one to form a minimal pair by either (a) using the grammatical counterpart in the original article, as in (4), or (b) constructing the grammatical sentence by focusing on the intended contrast of the example if we were unable to identify the grammatical sentence in the article, as in (5). Finally, 247 pairs consist of both sentences from the original articles, and in 90 pairs, the grammatical sentence was constructed by us using the above-mentioned criteria, resulting in a total of 337 × 2 = 674 sentences. For all the constructed sentences, two authors (well-trained syntacticians and native speakers of Mandarin) double-checked their grammaticality and whether they conveyed the intended contrast, as exemplified in the original articles. Note that all ungrammatical sentences are from the original authors.

4 Grammatical counterpart extracted from the original article (Wang and Liu, 2014, Example 29b and 29a)

4a	Text	*tā	yǒu	huí	xuéxiào.	
	Gloss	3sg	have	return	school	
	Trnsltn.	‘He has returned to school.’				
4b	Text	tā	méi	yǒu	huí	xuéxiào.
	Gloss	3sg	not	have	return	school
	Trnsltn.	‘He has not returned to school.’				

5 Grammatical counterpart (i.e., sentence (b)) created by the authors (Zhou and Jiang, 2014, Example 12a)

5a	Text	*tā	xǐhuān	zhangsn,	wǒ	bú	shì.
	Gloss	3sg	like	John,	I	not	COP
	Trnsltn.	‘He likes John. I am not.’					
5b	Text	tā	xǐhuān	zhāngsān	wǒ	bù	xǐhuān.
	Gloss	3sg	like	John,	I	not	like
	Trnsltn.	‘He likes John. I don’t.’					

Note that examples made of phrases (n = 73) were expanded into semantically neutral sentences with the aim that the sentential context should not influence the acceptability judgments of the phrases, as shown below.

6		The original example with only one noun phrase (a) vs. The expanded sentence used in our experiments (b) (Liao and Wang, 2011, Example 1b)				
6a	Text	liǎng	zhī	gou.		
	Gloss	two	CL	dog		
	Trnsltn.	'two dogs.'				
6b	Text	zhèlǐ	yǒu	liǎng	zhī	gou.
	Gloss	here	have	two	CL	dog
	Trnsltn.	'There are two dogs.'				

Five randomly sampled pairs from the 337 minimal pairs are presented in the Appendix. All the 337 pairs, data and analysis scripts are available at https://osf.io/z5pts.

2.2. Participants

To study the effects of dialectal/language background on acceptability judgments, we recruited two groups of participants: native speakers of Mandarin, born and raised in Beijing, where speakers are usually considered to speak Standard Mandarin, and bilingual speakers in Guangzhou (Canton), whose native languages are Cantonese and Standard Mandarin. To be eligible for this study, participants cannot have lived outside their respective region (Beijing/Guangzhou) for more than 2 years before the age of 18 years. A total of 489 participants were recruited for Experiment 1. Among them, 223 were from Beijing (N female = 142, median age = 20 years) and 266 were from Guangzhou (N female = 149, median age = 25 years).

The Guangzhou participants self-reported their proficiency on a scale of 10 in both Cantonese (mean = 8.5) and Mandarin (mean = 8). All participants were also asked whether they speak any other Chinese dialects/languages. Among the participants from Beijing, only one self-reported speaking a southern dialect (Wu) and three self-reported speaking a northern dialect. Among the participants from Guangzhou, 15 of them reported that they speak Hakka.

2.3. Procedure

Using online questionnaires administered by Qualtrics, Experiment 1 asked participants to rate the naturalness of sentences on a 7-point LS (displayed horizontally), with 1 on the left end labeled "very unnatural" and 7 on the right end labeled "very natural." The 674 sentences were randomly divided into six lists (each list contained approximately 110 sentences), where the two grammatically contrasting sentences from each pair were in different lists.

Each participant first rated the sentences from one of the six lists, with one sentence per page. We added two catch trials to each list and the participants had to answer both trials correctly for their data to be included in our analysis. Following Chen et al. (2020), in the catch trials, participants were asked to choose a specific rating (e.g., the number "5"), as opposed to rating the naturalness of the sentence. After completing the experiment, each participant was asked to fill out a comprehensive demographic and language background questionnaire. The questions were related to their age, gender, educational background, and which district in Beijing/Guangzhou they spent the most time in before the age of 18 years, among other such aspects.

The Guangzhou participants were asked to rate their proficiency in both Cantonese and Mandarin on a 1–10 LS. We believe that this can serve as a proxy for the language ability in both languages. The average self-reported rating for their Cantonese and Mandarin proficiency was 8.5 and 8, respectively.

The experiment took approximately 15 min to complete and each participant received 15 RMB for their participation.

2.4. Definition of replicated judgments

Following previous literature (Chen et al., 2020; Sprouse et al., 2013), judgments for a minimal pair are considered replicated in the LS rating task (Experiment 1) if and only if the ratings for the grammatical sentence in the pair are significantly higher than those of the ungrammatical one in the pair, using a two-tailed *t*-test.

2.5. Coding and analysis

To address the first two research questions, namely, whether informal judgments can be replicated in experimental contexts and whether ratings from participants with different language backgrounds would differ, we coded each

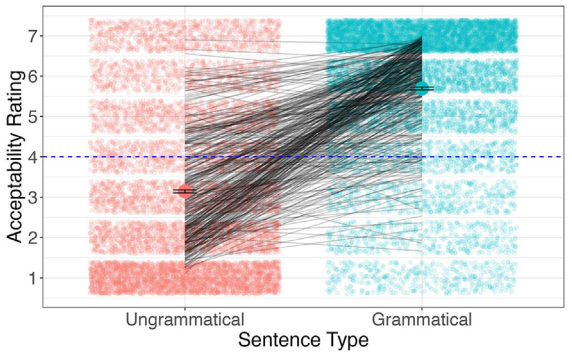

Fig. 1. Acceptability ratings of **Beijing participants** in Experiment 1: small circles represent individual data points. Black lines show the mean rating difference within each PAIR contrast across the two sentence types (ungrammatical vs. grammatical). Mean ratings for the two sentence types are highlighted with error bars representing the 95% confidence interval (CI).

sentence's grammaticality as reported by the original paper and the participants' region (Beijing vs. Guangzhou) after data collection. To determine which other factors may influence acceptability ratings, we included *sentence length*, measured by the number of characters, and *article language* (i.e., the language in which the journal article was written) as task-related factors. Additional subject-related factors, such as the *first author's origin*, the *participant's age*, and the *participant's gender*, were also included. The complete list of variables included in the data analysis is as follows:

- *Grammaticality*: A binary categorical variable that codes the original grammaticality of the sentence reported in the journal article. Levels: Grammatical, Ungrammatical.
- *Region*: A binary categorical variable that codes whether the participant is from Beijing or Guangzhou. Levels: Beijing, Guangzhou.
- *Article language*: A categorical variable that codes whether the article was written in English or Chinese. Levels: Chinese, English.
- *Sentence length*: A continuous variable that calculates the number of characters in each sentence.
- *First author's origin*: A binary categorical variable that codes the first author's origin.[4] Levels: Mainland, Non-mainland.
- *Age*: A continuous variable that records the age for each participant.
- *Gender*: A categorical variable that codes the participant's self-reported gender. Levels: Male, Female.
- *Education*: A categorical variable that codes the participant's self-reported education. Levels: Below undergrad, Undergrad, and Master.

The analyses were conducted using *R* version 4.0.5 (R Core Team, 2021). A mixed-effects logistic regression was run using *lme4* package version 1.1–27.1 (Bates et al., 2015), and plots were created using *ggplot2* package version 3.3.5 (Wickham, 2011).

To explore the effect of the Cantonese-Mandarin bilingual status on acceptability judgments, we used two levels of analysis. The first is what we call the "overall" analysis, where `ratings` for all pairs are used as the dependent variable and `region` is used as one of the independent variables in the regression model. We then examined whether `region` is a significant main factor in the model. The second analysis is "pair-wise" in nature. This analysis checks, for each minimal pair, whether the replication status is the same in both regions, where replication status is defined as whether the judgments of participants in that region are the same as the judgments in the published articles. For Experiment 1,

[4] We operationalized this as where the first author lived before the age of 18 years. This was determined by looking up each author's personal webpage and determining their educational background. For instance, if the author grew up in Beijing and went to college there, but completed their graduate studies in the US, this person would be categorized as having a Chinese mainland origin. Ultimately, 31 authors were identified as coming from the Chinese mainland, 24 from Taiwan, 4 from Hong Kong, 3 from Singapore, 1 from the USA, 1 from Japan, and 1 from Germany. To avoid categories with too few data points, we collapsed all the non-Chinese-mainland authors into one category.

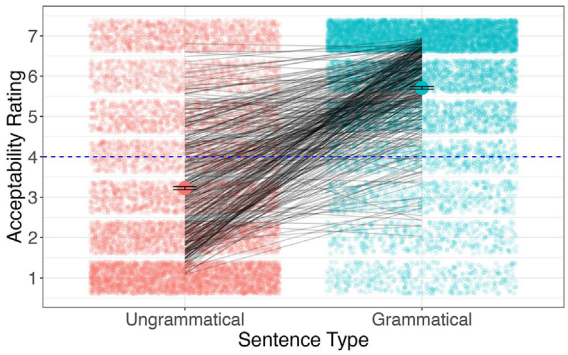

Fig. 2. Acceptability ratings of **Guangzhou participants** in Experiment 1: small circles represent individual data points. Black lines show the mean rating difference within each PAIR contrast across the two sentence types (grammatical vs. ungrammatical). Mean ratings for the two sentence types are highlighted with error bars representing the 95% CI.

this is operationalized as whether the good sentence has a significantly higher rating than the bad sentence, using a two-tailed *t*-test, as discussed above.

2.6. Results for Experiment 1

Before the statistical analysis was conducted, we excluded data points from those participants who incorrectly answered the catch trials or lived outside Beijing/Guangzhou for more than 2 years before the age of 18 years. In total, data from 187 participants from Beijing and 191 participants from Guangzhou were included in the final analysis.[5]

As can be seen from Figs. 1 and 2, overall, the mean acceptability for the two sentence types (grammatical vs. ungrammatical) differs notably, regardless of which region the participants were from, with the "grammatical" sentences being rated generally higher (mean for Beijing participants = 5.69, mean for Guangzhou participants = 5.71) than the "ungrammatical" ones (mean for Guangzhou participants = 3.23, mean for Beijing participants = 3.14). This suggests an overall replication of grammaticality judgments given in the published articles sampled.

2.6.1. Modeling of all potential predictors

A linear mixed-effects regression model was constructed to predict the participants' ratings (z-scored) of each sentence, with the following fixed effects: *grammaticality*, *region*, *sentence length*, *article language*, *first author's region*, *gender*, *education*, and *age*. *Grammaticality* by *pair* was included as the random slope.[6] *Length* was centered and scaled using the *scale()* function in R. *Age* was centered by subtracting the age of each participant from that of the youngest participant so that the model coefficients represent whether the older participants are systematically different from their younger counterparts. All categorical variables were sum-coded, allowing us to compare responses under specific experimental conditions using the grand mean.

The model output, as shown in Table 3, revealed a significant effect of *grammaticality*, suggesting that the grammatical sentences were rated significantly higher compared with the grand mean (β = 1.0129e + 00, $p < 0.001$). *Region* was not significant, indicating that the participants from Beijing did not differ significantly from the grand mean in terms of their acceptability ratings (β = 5.236e-05, p = 0.99). The effect of *paper language* was significant. That is, ratings for journal articles written in English were lower than those written in Chinese (β = −2.028e-01, $p < 0.001$). The effect of *sentence length* was significant (β = −4.084e-02, p = 0.04), suggesting that the acceptability rating tended to decrease significantly for longer sentences (as measured by the number of characters). Other variables such as *gender*, *age*, and *first author's region* did not show significant effects.

To further investigate whether the language proficiency of participants would influence the way they rated the sentences, we conducted a similar statistical analysis, focusing only on the ratings from the Guangzhou participants. In addition to the predictors of interest mentioned above, we included participants' self-reported Mandarin/Cantonese

[5] A few of our Guangzhou participants self-reported that they know some other dialects, such as Hakka. However, these were not removed because this group did not behave as outliers.

[6] Treating *grammaticality* by *participant* as a random effect in the model produced a zero-variance estimate; therefore, it was excluded from the final model.

Table 3
Model results: Rating ~ Grammaticality + Region + Sentence Length + Article language + Education + Age + Gender + First-author region + (Grammaticality | Pair)

	Estimate	Std. Error	df	t value	Pr(> | t |)
Intercept	3.766e-02	2.391e-02	9.128e + 02	1.58	0.12
Grammaticality (Gram.)	1.129e + 00	4.019e-02	6.428e + 02	28.090	< 0.001 ***
Region (Beijing)	5.236e-05	6.398e-03	4.180e + 04	0.01	0.99
Sentence Length	−4.084e-02	1.983e-02	6.385e + 02	−2.06	0.04 *
Article language (English)	−2.028e-01	4.966e-02	6.426e + 02	−4.09	< 0.001 ***
Education (BelowUndergrad)	1.562e-04	1.343e-02	4.181e + 04	0.01	0.99
Age	1.778e-05	5.595e-04	4.185e + 04	0.03	0.97
Gender (Female)	−4.531e-04	7.467e-03	4.187e + 04	−0.06	0.95
First author's region (Mainland)	−4.572e-02	4.573e-02	6.422e + 02	−1.04	0.30

proficiency scores in the model. The participants' Mandarin and Cantonese proficiency scores were highly correlated, suggesting that the Guangzhou participants were bilingual and proficient in both Mandarin and Cantonese. We found that participants' self-reported Mandarin (or Cantonese) score did not significantly influence their ratings of sentences (β = −5.979e-06, p = 0.99). Note that these proficiency data were only collected in Experiment 1 and only among Guangzhou participants.

2.6.2. Replication of grammaticality

To check whether the judgments from our rating experiment converge with the judgments in the published articles, we collected the ratings for each minimal pair and checked 1) whether the grammatical sentence was rated as more acceptable than the ungrammatical one (*z*-scored); and 2) whether the difference between the ratings of the two sentences was statistically significant, using *t*-tests. For the *t*-tests, none of the predictors (i.e., *sentence length, article language, education, age, gender, and first author's region*) from the previous model were included because they were not relevant for this particular analysis.

There are four possible replicability outcomes for each pair, as listed in Table 4. Following previous literature (Sprouse and Almeida, 2012; Chen et al., 2020), only pairs in the "replicated" category, that is, pairs where the grammatical sentence is rated significantly higher than its ungrammatical counterpart in terms of acceptability, are considered successfully replicated; specifically, the experimental and introspective judgments have converged. Based on this criterion, of the 337 pairs in Experiment 1, 289 pairs replicated the judgments from the journal articles among Beijing participants (replication rate = 85.8%), while 291 pairs replicated the judgments among Guangzhou participants (replication rate = 86.4%), as shown in Table 4. Among the unreplicated pairs, 14 pairs were not replicated by the Beijing participants but replicated by the Guangzhou participants, another 12 pairs were not replicated by the Guangzhou participants but replicated by the Beijing participants, and 34 pairs were replicated by neither the Beijing nor the Guangzhou participants (see Fig. 4 for details). Fig. 3 shows example pairs for the four possible replicability outcomes in Experiment 1 discussed above.

2.6.3. Differences between Beijing and Guangzhou participants

In Experiment 1, the overall analysis (cf. Table 3) shows that the predictor `region` does not reach statistical significance (p = 0.99). This suggests that, when examining all the ratings for all pairs together, there is no significant difference between the two regions.

Using pair-wise analysis, we further observed that participants from both regions converged on what counts as an acceptable sentence for a given pair in 311 of the 337 pairs (see Fig. 4). Specifically, both participant groups *agreed* with the judgments from the journal articles in 277 out of 337 pairs (replicated in both regions), while they both *disagreed* with

Table 4
Number of pairs for each replicability outcome; only the "replicated" outcome is considered replicated experimentally.

Replicability outcome	Numerical rating direction	Sig.	N Beijing	N Guangzhou
Replicated	Gram > ungram	Sig	289	291
Not replicated	Gram > ungram	Non-sig	28	26
Not replicated	Gram < ungram	Non-sig	16	16
Not replicated	Gram < ungram	Sig	4	4

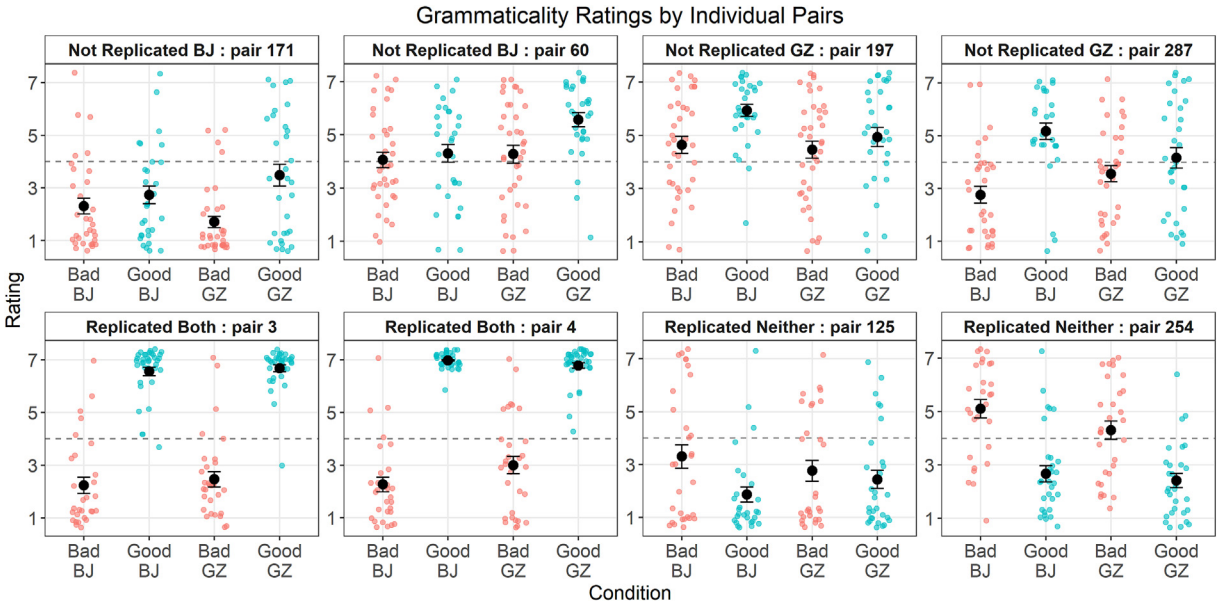

Fig. 3. Example pairs for four groups in Experiment 1: (1) Not replicated in Beijing (BJ), (2) not replicated in Guangzhou (GZ), (3) replicated in both regions, and (4) replicated in neither region. Small circles represent individual data points. The color red represents "ungrammatical, bad" sentences and the color blue denotes "grammatical, good" ones.

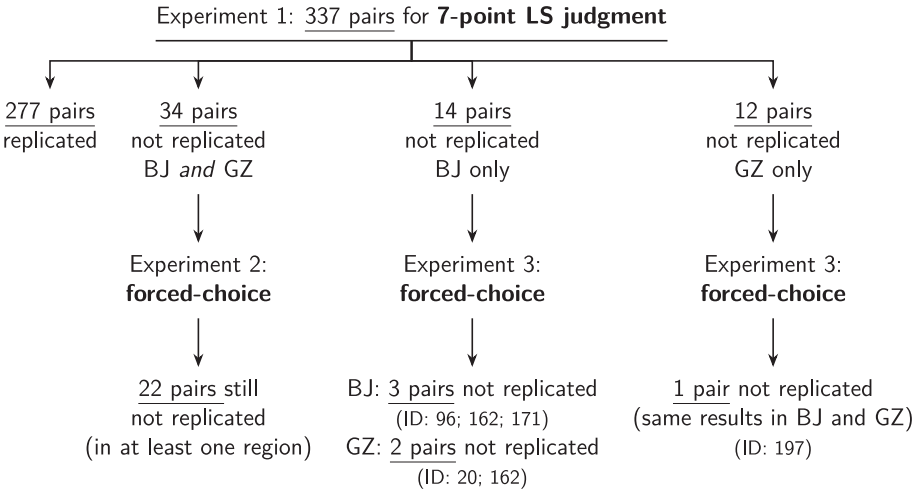

Fig. 4. Summary of results for all three experiments. "Replicated" means that the judgments of our participants are in line with the linguists' judgments given in the journal articles the examples were sampled from. BJ: Beijing, GZ: Guangzhou.

linguists' judgments in 34 out of 337 pairs (not replicated in both regions). This resulted in 26 pairs with different judgments in the two regions. Thus, we conclude from Experiment 1, using an LS rating task and examining each pair individually, **that 26/337 = 7.7% of the pairs receive different judgments from the two regions.** This suggests that, overall, Cantonese and Mandarin bilinguals share the same Mandarin grammar but may differ in judgments for a small percentage of pairs.

2.7. Prelude to Experiments 2 and 3

To further examine whether the judgments of the unreplicated pairs were indeed non-replicable, all unreplicated pairs were tested in Experiments 2 and 3 using an FC task, where participants must choose the more natural sentence from a

given minimal pair. Switching from an LS rating task to an FC task allows us to see, in a more direct way, how participants would rate each sentence against its minimally different counterpart in a given contrast, which is a common practice to elicit syntactic judgments in theoretical syntax. The idea is to see whether the pairs that failed to replicate in an LS rating task could be replicated in an FC task, as previous literature has indicated that FC tasks are more sensitive to categorical decisions than LS tasks (Sprouse, 2018). Further, an FC task would provide another angle to investigate any regional differences in acceptability judgment.

To better separate the different types of unreplicated pairs, the 34 pairs that were not replicated in *both* regions were tested in Experiment 2. The 26 (12 + 14) pairs that were not replicated in *only one* region were tested in Experiment 3. Fig. 4 presents a flowchart and summary of these experiments.

3. EXPERIMENT 2

In contrast to Experiment 1, in which participants rated sentences one at a time to determine the relative location of each sentence on the acceptability spectrum, Experiment 2 used an FC task in which participants were presented with a minimal pair and had to decide which of the sentences was more natural.

3.1. Stimuli

The minimal pairs replicated neither in Beijing nor in Guangzhou from Experiment 1 were used as test items in Experiment 2 (N = 34 pairs). A total of 17 fully replicated pairs from Experiment 1 were used as control items to determine whether replicated judgments would also be rated as expected in an FC task and to serve as filler items. Taken together, a total of 51 pairs were tested in Experiment 2.

3.2. Participants

The participants in Experiment 2 were recruited from both regions (Beijing and Guangzhou), as in Experiment 1. For Experiment 2, we recruited 40 participants (F = 32, M = 8, median age = 19 years) and 38 participants (F = 36, M = 2, median age = 20 years) from Beijing and Guangzhou, respectively. None of these participants were involved in Experiment 1, and they were paid 10 RMB after the experiment. Fewer participants were recruited for Experiments 2 and 3 because these two experiments included significantly fewer items than Experiment 1. Thus, in order to collect approximately 30–40 judgments for each experimental item, as in Experiment 1, fewer participants were needed.

3.3. Procedure

As in Experiment 1, Experiment 2 was conducted online, using Qualtrics. Participants from both regions were instructed to choose the more natural sentence from a minimal pair, one pair per page. All pairs were randomized. Two catch trials were again included to check the participants' attention. Both catch trials had to be answered correctly for a participant's data to be included in the analysis. In Experiment 2, for each catch trial, participants were asked to select a specific sentence that was explicitly requested.

3.4. Definition of replicated judgments

For the FC task adopted in Experiment 2, a minimal pair is considered replicated if and only if the informal acceptability judgment in the published article turns out to be a significant predictor, in a logistic regression, of the binary choice regarding which sentence in the pair is the better of the two. Specifically, we follow Chen et al. (2020) and use `glm(Choice~1,family=binomial(link="logit"))` for each region separately, to see if participants choose significantly more good sentences than bad ones.

3.5. Results for Experiment 2

Prior to data analysis, we verified that all participants answered the two catch trials correctly; thus, no data points were excluded. As in Experiment 1, we first report the general pattern based on the overall analysis, followed by an analysis targeting comparisons of individual pairs across the two regions.

3.5.1. General pattern and regional differences

Fig. 5 shows the overall pattern of results from Experiment 2 for both regions. For the control and test groups, both Beijing and Guangzhou participants preferred "good" sentences to "bad" ones. The good vs. bad contrast is starker for control group items compared to test group items. That is, in the control group, "good" sentences were almost always selected. However, the pattern is less extreme for the test group pairs. Yet, in general, Beijing participants appear to choose the "good" sentences more often than participants from Guangzhou.

To confirm these patterns statistically, we conducted a mixed-effects logistic regression model to predict the proportion of choosing a "good" sentence as the better one for all the pair contrasts using *group* (control vs. test) and *region* (Beijing vs. Guangzhou) as fixed effects in a two-way interaction, and *group* by *participant* and *group* by *pair* as random slopes. The inclusion of the *group*region* interaction allowed for a better examination of whether in the FC context, participants from different regions would exhibit differences in their judgments for the two groups of items. All the categorical predictors were sum-coded. Model results, as illustrated in Table 5, reveal that overall, the proportion of choosing a "good" sentence as the better one was significantly higher for the control group pairs compared to the grand mean (β = 1.74, $p < 0.001$). In addition, a marginally significant effect of *region* was noted: Beijing participants tended to select more "good" sentences (β = 0.16, p = 0.07). This shows that the distinction between good and bad sentences is clearer for Beijing participants than for their Guangzhou counterparts. The interaction between *group* and *region* was not significant (β = 0.03, p = 0.72).

To better understand whether there were differences in judgments between participants from different regions for different individual Group pairs, post hoc pairwise comparisons were extracted from the fit model using *emmeans*, as illustrated in Table 6. The most indicative result is that for test items, Beijing participants chose a significantly higher proportion of "good" sentences (β = 0.26, $p < 0.02$). This suggests that the Beijing participants were more aligned with judgments from the journals for items in the test group. For items in the control group, Beijing and Guangzhou participants behaved similarly (β = 0.39, p = 0.24), as participants from both regions significantly preferred "good" sentences.

3.5.2. Replicability of individual pairs across regions

To examine differences in the judgments of participants from the two regions for every individual pair in the test group, we fit logistic mixed-effects models to each of the contrasts for Beijing and Guangzhou participants. As illustrated in Fig. 6, the highlighted pairs are those that were not replicated; that is, the participants' judgments were different from those in the journal articles. Nineteen pairs were not replicated among the Beijing participants, while there were 20 such pairs among the Guangzhou participants. The union of the two sets shows that in total, 22 out of 34 pairs were not replicated in at least one region in this FC experiment. We discuss the categorization of these pairs in Section 5.4.

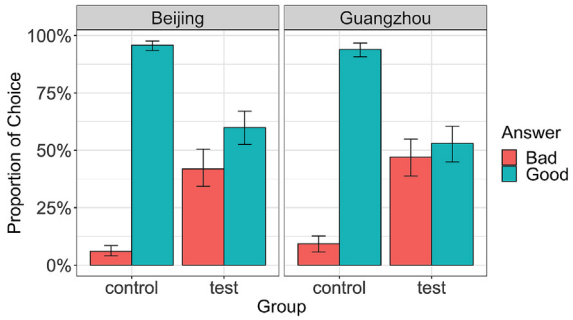

Fig. 5. Overall pattern in Experiment 2. Bars represent the proportion of choosing a preferred sentence from a contrast. Error bars indicate the 95% CI.

Table 5
Model results of Experiment 2: Answer ~ Group × Region + (1 + Group|Participant) + (1 + Group|Pair).

Fixed effects	Estimate	Std.Error	z-value	Pr(>|z|)
Intercept	1.99	0.21	9.46	<0.001
Group (Control)	1.74	0.21	8.33	<0.001
Region (Beijing)	0.16	0.09	1.81	0.07
Group (Control): Region (Beijing)	0.03	0.08	0.36	0.72

Table 6
Post-hoc pairwise comparisons in Experiment 2.

Condition	Contrast	Estimate	Std.Error	z-value	Pr(>|z|)
Control	Beijing - Guangzhou	0.39	0.33	1.75	0.24
Test	Beijing - Guangzhou	0.26	0.12	2.25	0.02

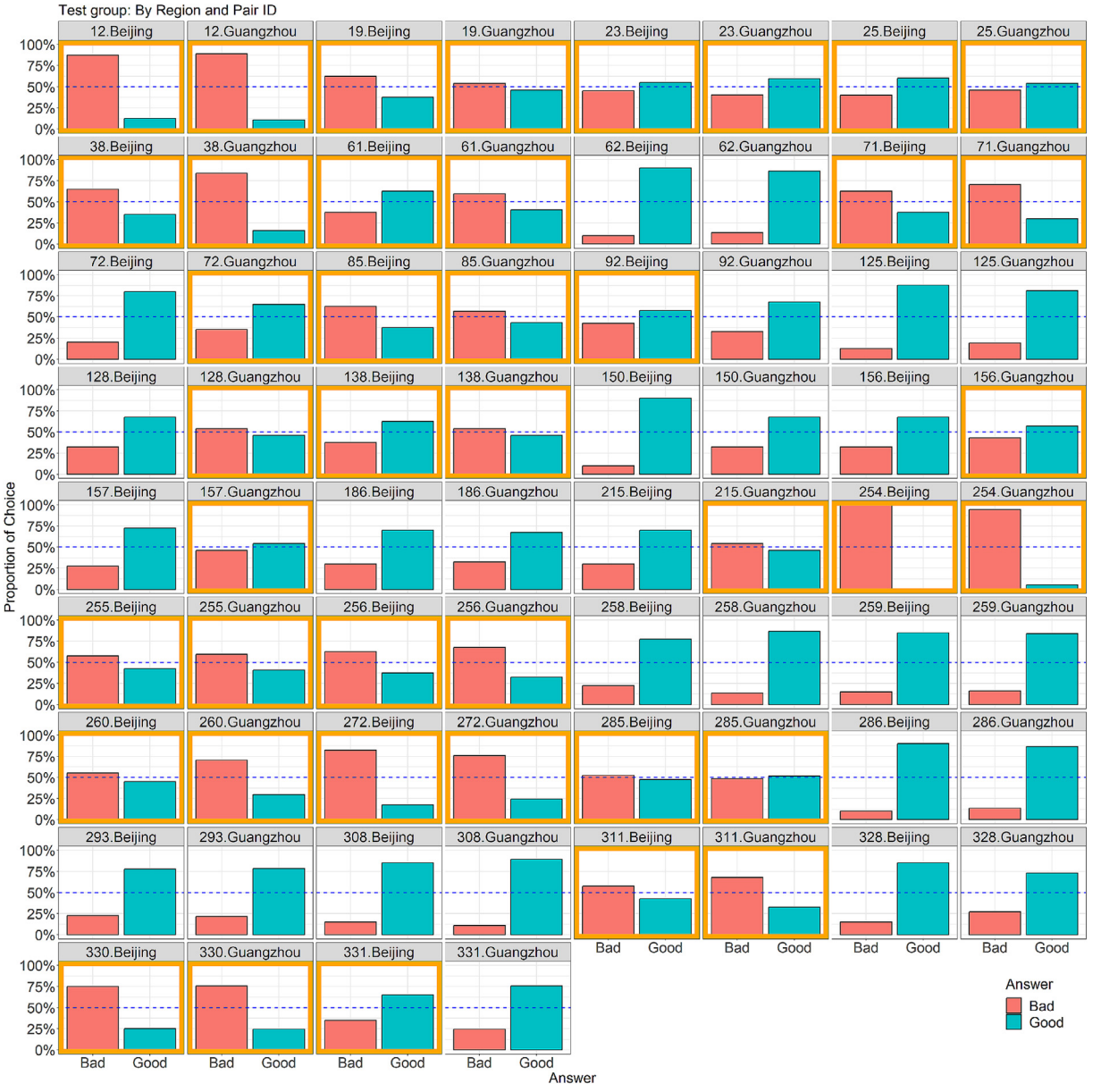

Fig. 6. Patterns of individual pairs in the test group of Experiment 2: The red bar represents the number of participants choosing the bad sentence, while the blue bar represents the number choosing the good sentence. An orange outline indicates that the contrast is not replicated, that is, the good sentence was not chosen significantly more. The numbers in the facets refer to the pair IDs and the region.

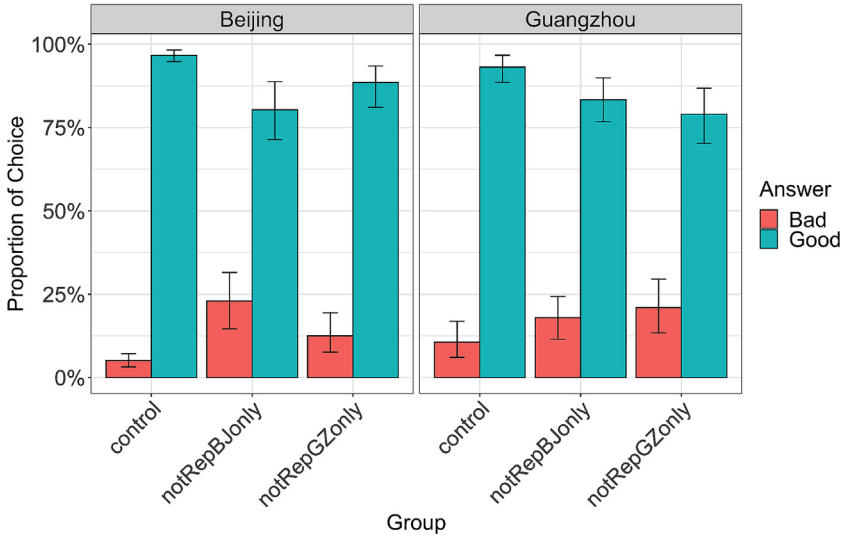

Fig. 7. Overall pattern in Experiment 3. Bars represent the proportion of choosing a preferred sentence from a contrast. Error bars indicate the 95% CI.

Notably, several pairs have a different replication status in the two regions. That is, pairs 92 and 331 are replicated in Beijing but not in Guangzhou. Another five pairs are replicated in Guangzhou, but not in Beijing: pairs 72, 128, 156, 157, and 215. These account for 20.6% (7/34) of the test pairs in Experiment 2 and 2% (7/337) of all pairs in our study. For these pairs, the LS task in Experiment 1 did not show any difference between the participants in Beijing and Guangzhou, whereas the FC task does. We discuss this further in the Discussion section.

4. EXPERIMENT 3

The goals of Experiment 3 are twofold: 1) to examine whether the pairs not replicated in one region using LS ratings would remain unreplicated with the FC paradigm in the *same* region, and 2) to examine whether those pairs would remain replicated in the *other* region. That is, under the FC paradigm, if the pairs that were not replicated by the Beijing participants are still unreplicated in Beijing but replicated in Guangzhou, then this would be further evidence that the Beijing participants alone do not agree with the original judgments of these pairs.

4.1. Stimuli

As in Experiment 2, we take a subset of pairs from Experiment 1 for an FC task. While Experiment 2 tested the pairs that were not replicated in either region, in Experiment 3, we examine the pairs that were not replicated in *only* one region. That is, we used the 14 pairs that were not replicated from our Beijing participants (notRepBJonly) and the 12 pairs unreplicated in Guangzhou (notRepGZonly) as the test items (see Fig. 4). We used the same control items as in Experiment 2.

4.2. Participants

A total of 37 participants from Beijing (F = 31, M = 6, median age = 21 years) and 49 participants from Guangzhou (F = 39, M = 10, median age = 23 years) were recruited for Experiment 3. None of these participants were involved in Experiment 1 or 2. They were paid 10 RMB for their participation.

4.3. Procedure and definition of replicated

Experiment 3 was conducted online using Qualtrics and followed the same procedure as Experiment 2. The definition for a pair to be considered “replicated” is the same as in Experiment 2.

Table 7
Statistical analysis results of Experiment 3: Answer ~ Group × Region + (1 + Group|Participant) + (1 + Group|Pair).

Fixed effects	Estimate	Std.Error	z-value	Pr(>|z|)
Intercept	2.58	0.20	13.18	<0.001
Group (Control)	1.32	0.29	4.56	<0.001
Group (notRepBJonly)	−0.72	0.26	−2.80	0.01
Region (Beijing)	0.24	0.09	2.80	0.01
Group (Control): Region (Beijing)	0.19	0.10	1.86	0.06
Group (notRepBJonly): Region (Beijing)	−0.36	0.08	−4.45	<0.001

4.4. Results for Experiment 3

4.4.1. General pattern

As in Experiment 2, all participants answered the catch trials correctly and no data points were excluded. Fig. 7 shows the overall pattern of how Beijing and Guangzhou participants judged the pairs in the three groups of sentences in Experiment 3: control, notRepBJonly, and notRepGZonly. In the control condition, "good" sentences were always selected as the better ones. For pairs that were replicated in only one of the two regions, participants demonstrated more uncertainty toward pairs that were originally not replicated among participants from the same region. In other words, even though both Beijing and Guangzhou participants selected reliably more "good" sentences for pairs that were originally replicated either in Beijing or Guangzhou, the tendency for Beijing participants to select "bad" sentences was higher for pairs that were not originally replicated in Beijing. Similar patterns were also found for Guangzhou participants.

These patterns are further confirmed by statistical modeling (see the model output in Table 7). A mixed-effects regression model was configured to predict participants' choices for each pair, with *group* (i.e., sentence groups) and *region* (i.e., participants' region) in a two-way interaction as fixed effects, allowing us to statistically test whether Beijing and Guangzhou participants judged different groups of sentences differently. *Group* by *participant* and *group* by *pair* were included as random slopes. Overall, the portion of "good" sentences selected was significantly higher for items in the control group (β = 1.32, $p < 0.001$). In addition, participants selected a significantly lower proportion of "good" sentences for the notRepBJonly group (β = −0.72, p = 0.01). This tendency is even more pronounced for Beijing participants, as illustrated by the interaction between *group* (notRepBJonly) and *region* (Beijing) (β = −0.36, $p < 0.001$). Moreover, there was a main effect of *region*, suggesting that participants from Beijing tended to choose "good" sentences as the better one for each given pair (β = 0.24, p = 0.01). However, the interaction between the control group and Beijing region is only marginally significant (β = 0.19, p = 0.06), implying that for the control items, Beijing and Guangzhou participants were consistent in their choices, though the choices were more categorical for Beijing participants at the group level. This result indicates that across the population, the choices between grammatical and ungrammatical sentences were more distinct than those made by Cantonese-Mandarin bilinguals.

Post-hoc comparisons were extracted using *emmeans*. These results further suggest that there were differences between Beijing and Guangzhou participants, but that the differences were only significant for items in the notRepGZonly group($\beta = 0.82, p < 0.001$), as Beijing participants were more likely to choose the "good" sentences, thus making them more aligned with the original judgments. A complete list of post hoc comparisons can be found in the Appendix.

4.4.2. Regional differences in individual pairs

Following the procedure of Experiment 2, we fit logistic mixed-effects models to each of the contrasts in Experiment 3 for Beijing and Guangzhou participants (see Fig. 8). Of the 12 notRepGZonly pairs, only one (pair 197) remained unreplicated in both regions. As for the 14 notRepBJonly pairs, three pairs were not replicated for Beijing participants (pairs 96, 162, 171) and two pairs were not replicated for Guangzhou participants (pairs 20, 162). We consider a pair to be replicated only when it is replicated in *both* regions. Thus, after this experiment, five pairs were still not replicated. We discuss the categorization of these unreplicated pairs in Section 5.4.

We can also categorize all test pairs (notRepBJonly + notRepGZonly) based on whether there is a difference in the replication status between Beijing and Guangzhou, which results in the following two groups.

Group 1: Qualitative difference between Beijing and Guangzhou. (N = 3). These are pairs where Beijing and Guangzhou participants disagree on whether the minimal pair forms a contrast. Following Chen et al. (2020), we fit a logistic mixed-effects model to the data of each minimal pair per region, and the results show that **only three out of the 43 pairs belong to this group** (pairs 20, 96, and 171)).[7]

[7] Specifically, we used the following model `glm(Choice~1,family = binomial(link = "logit"))` for each region separately to see if participants chose significantly more good sentences than bad ones, and to find the pairs where the two regions disagree.

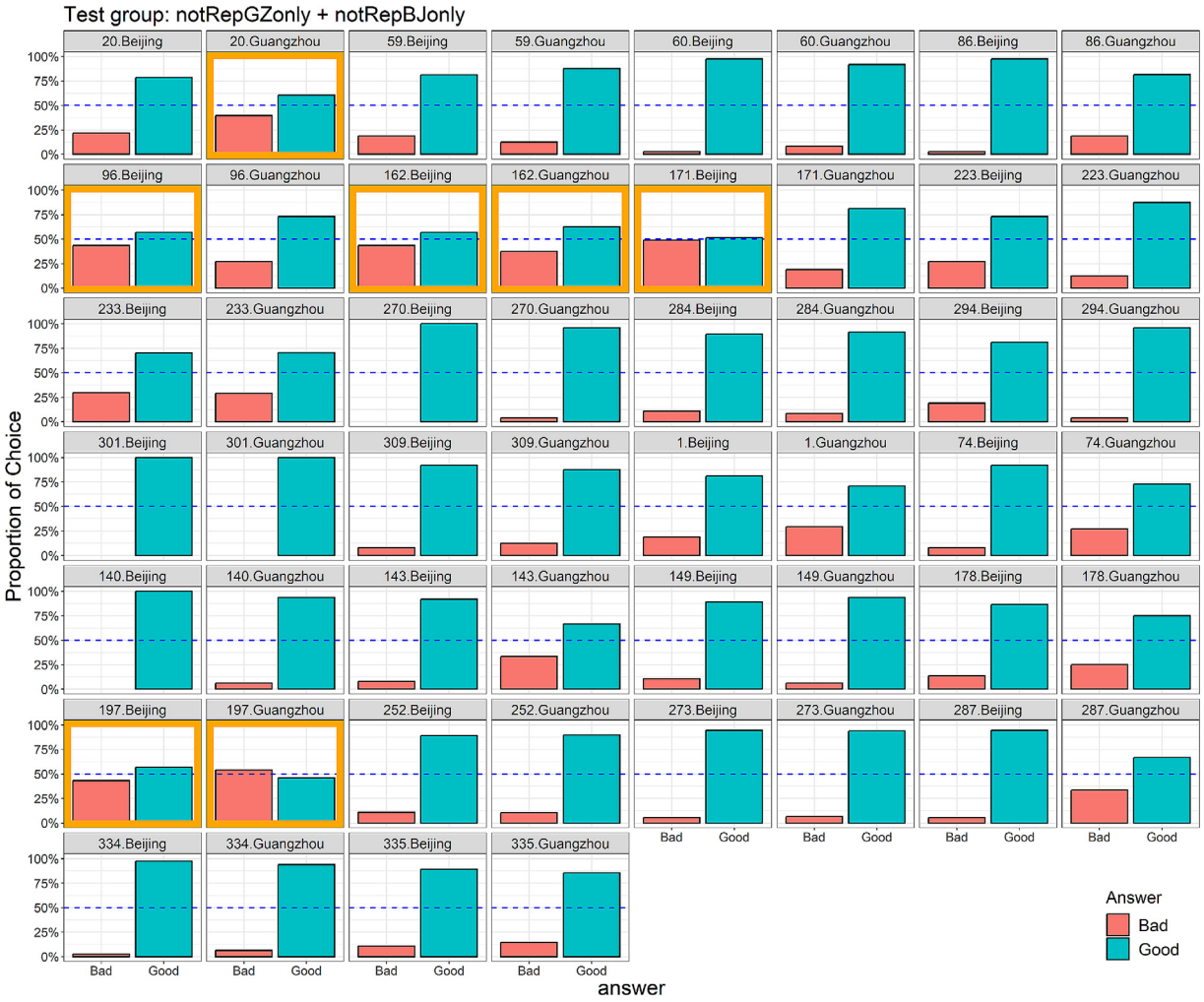

Fig. 8. Patterns of individual pairs in the test group of Experiment 2: The red bar represents the number of participants choosing the bad sentence, while the blue bar represents the number choosing the good sentence. An orange outline indicates that the contrast is not replicated, that is, the good sentence was not chosen significantly more.

Group 2: No difference between the two regions. All other 36 pairs show no difference between the two groups of participants.[8]

Combining the findings from the overall model presented in Table 7 and the categorization into the two groups mentioned above, we conclude as follows. For minimal pairs in Experiment 3, (1) Beijing participants selected the "good" sentences more often than Guangzhou participants, and (2) for three pairs, the difference between participants from the two regions is large enough to qualify as a distinction in the binary acceptability judgment. These three pairs account for 6.9% (3/43) of the pairs in Experiment 3 and 0.9% (3/337) of all pairs in the entire sample.

In other words, **although Guangzhou participants demonstrate a less sharp distinction at the group level between the good sentence and the bad one for a minimal pair, they still show a clear preference for the former.**

With these results in mind, we now discuss the main findings and implications of this study.

[8] Neither BJ nor GZ participants agree with linguists' judgments (N = 2, pair 197 and 162); both BJ and GZ participants agree with linguists (N = 34)

5. DISCUSSION

This study investigated the reliability of acceptability judgments in journal articles, focusing on Mandarin Chinese. Aiming to present a representative sample, we conducted three experiments using 337 minimal pairs randomly sampled from 10 academic journals on Chinese syntax (broadly defined). Experiment 1 used a 7-point LS rating task. Unreplicated pairs from Experiment 1 were further examined in Experiments 2 and 3, using an FC task. Our participants came from two distinctive language backgrounds—native speakers of Standard Mandarin born and raised in Beijing, and bilingual speakers of Cantonese and Standard Mandarin born and raised in Guangzhou—allowing us to empirically examine the effect of language background on acceptability judgments.

5.1. Replicability of acceptability judgments for Chinese in journal articles

The results of Experiment 1 showed convergence rates of 85.8% (289/337) and 86.4% (291/337) for the Beijing and Guangzhou participants, respectively (see Section 2.6.2). The convergence rate refers to the percentage of pairs that receive the same judgments in our experiments as the judgments given in the journal articles. We consider only the pairs that are replicated in *both* regions in these three experiments as fully replicated, which leaves us with 27 pairs still not replicated after all three experiments (see Fig. 4). Thus, the final convergence rate is (337–27)/337 = 92%. The different convergence rates between the LS rating task (Experiment 1) and the FC task (Experiments 2 and 3) further demonstrate that FC tasks are more sensitive in capturing the differences between grammatical and ungrammatical sentences in pairwise contrasts, consistent with previous studies comparing different acceptability judgment tasks (Sprouse and Almeida, 2017; Chen et al., 2020). Therefore, utilizing different tasks is beneficial, as they complement each other and can provide a more holistic view when examining acceptability judgments.

Compared with previous studies on English, our convergence rates are similar but lower than those reported for journal articles. For example, Sprouse et al. (2013) report a 95% convergence rate for English sentences from *Linguistic Inquiry* articles. One possible reason for this is that our sentences come from a more diverse set of sources—10 journals with different editing and publishing standards and styles—than Sprouse et al. (2013), who focused on only one journal.

For Chinese, our convergence rates are also lower than those reported for sentences from a Chinese syntax textbook by Chen et al. (2020): 89.2% for LS rating tasks and 96.8% for FC tasks. This difference may stem from the nature of the sentences used; textbook sentences tend to be less controversial than those found in journal articles, where linguists engage in debates on a range of syntactic issues.

Taken together, our results suggest that the judgments in journal articles on Chinese syntax are reliable overall and that the convergence rate is comparable to that of previous studies with similar representative samples. It is worth mentioning that unlike Linzen and Oseki (2018), whose goal was to study controversial contrasts, we set out to collect a representative sample by sampling examples from a diverse set of 10 journals, to estimate the replicability of sentences from syntax research conducted by researchers from different Chinese-speaking communities in the last decade. Our replication rate is unsurprisingly higher than that reported by Linzen and Oseki (2018), who found that “half of the Hebrew contrasts and a third of the Japanese contrasts did not replicate in formal experiments” (pp. 16). We suggest that the contrasts that are not replicated in our study could serve as a starting point for future research seeking to identify controversial issues in Chinese syntax. Some of these pairs are discussed in Section 5.4, and all of them are available at https://osf.io/z5pts/.

5.2. Impact of language background on acceptability judgments

Regarding the impact of language background on judgments, we examine both the language background of our participants (Section 5.2.1) and that of the syntacticians who authored the examples (Section 5.2.2).

5.2.1. Language background of the participants

One of the primary objectives of this study was to investigate the potential variations in judgments between Mandarin-speaking participants from Beijing and Mandarin-Cantonese bilingual participants from Guangzhou. This inquiry stems from a critical reassessment of the ideology of native-speakerism, which often underpins linguistic research. Monolingual native speakers are often presumed to have superior access to language knowledge, which is commonly evaluated through grammaticality judgments. In the realm of Chinese language competence research, judgments from Beijing Mandarin speakers are often prioritized over those speaking other regional varieties or from multilingual individuals whose first language is a different Sinitic language.

As shown in Table 8, for each experiment, we investigated the differences between our Beijing and Guangzhou participants based on two levels. First, we examined the overall level, using the overall statistical model with all data points,

Table 8
Summary of differences between Beijing and Guangzhou participants in each experiment.

Analysis	Operationalization	Exp 1	Exp 2	Exp 3
Overall	Is *region* a significant predictor?	$p = 0.99$	$p < 0.07$	$p < 0.01$
Pair-wise	$\frac{\#\text{pairs with different replication status}}{\#\text{total pairs in exp}}$	$\frac{26}{337}$	$\frac{7}{51}$	$\frac{3}{43}$

and checked whether the predictor `region` was significant or not. Second, we examined the issue at the pair level, checking whether there is a difference in the replication status between the two regions (i.e., replicated in Beijing but not in Guangzhou, or vice versa).

In the overall-level analysis, we found that `region` is significant in Experiment 3 ($p < 0.01$), marginally significant in Experiment 2 ($p = 0.07$), and not significant in Experiment 1 ($p = 0.99$). This suggests that for the LS rating task, we do not observe an overall meaningful difference between Beijing and Guangzhou participants. The reason for the difference observed in Experiment 3 can be attributed to the fact that the test items in Experiment 3 are, in fact, those that displayed a regional difference in Experiment 1. Thus, this regional difference is more pronounced only when these items are tested in Experiment 3. It is also important to mention that in Experiment 3, the regional difference is that Beijing participants make a sharper distinction between acceptable sentences and those that are not, compared to those from Guangzhou.

For the pair-level analysis, we find that in Experiment 1, when the LS rating task was used, 26 pairs out of 337 pairs have a different replication status in the two regions (7.7% of the 337 pairs). For Experiments 2 and 3, which involved FC tasks, we find that seven pairs and three pairs show such a difference, respectively, equivalent to 3% of the 337 total pairs ((7 + 3)/337 = 3%). We take this to mean that for the 7.7% showing a different replication status with the LS method, only a small number of these still manifest a regional difference with the FC task.

Thus, using the two levels of analyses for comparison ("overall" and "pair-wise"), we tentatively conclude as follows: Speakers from the two different language backgrounds mostly make *qualitatively* identical acceptability judgments on our representative sample of Mandarin Chinese minimal pairs, even though, for a few pairs, when presented in an FC task, Beijing participants tend to have sharper (or more categorical) judgments than their Guangzhou counterparts.

The finding that bilingual perceptions of linguistic categories are less discrete or categorical has also been reported for the perception of phonological categories, where the gradient categorical boundaries may help bilinguals to "flexibly shift between languages" (Kutlu et al., 2022, pp.7).

This may also be the case for our Mandarin-Cantonese bilinguals, who are exposed to both languages daily and need to shift between them frequently. It is thus likely that, for them, the distinction between a good and bad sentence may be less categorical than for monolinguals who are mostly exposed to Mandarin only. It is worth re-emphasizing that, despite the more gradient nature of their judgments, when forced to make a binary decision, bilinguals from Guangzhou share the same judgments as monolingual speakers from Beijing, for the vast majority of our test items.

In the grammaticality judgment literature, our study differs from the Yale Grammaticality Project (Zanuttini et al., 2018) where clear dialectal differences have been found for various syntactic constructions. We found that the differences induced by bilingualism or dialects are much more subtle. We argue that this can be attributed to the fact that our experiment and the Yale Grammaticality Project have different goals. The Yale Grammaticality Project focuses on syntactic phenomena that are potentially judged differently by speakers of various varieties of English. Our study is not based on the assumption that the pairs that we sampled should exhibit cross-linguistic or dialectal differences. Instead, we focused on examining the degree of judgment variation in a representative sample, which can shed light on whether judgments from less "typical" native speakers, such as multilinguals, are qualitatively distinctive from those of the monolingual native speakers. In addition, while the Yale Grammaticality Project is targeted at dialectal grammar, in our study, we focus on how participants from Beijing and Guangzhou make judgments about a target language that they share, that is, the grammar of Standard Mandarin. While Cantonese and Mandarin are distinctive languages with different grammar, our study shows that Cantonese-Mandarin bilinguals make similar, if less categorical, judgments to those of the Beijing participants for Standard Mandarin grammar.

5.2.2. Language background of the syntacticians

Comments often heard in syntax classrooms or about syntactic research relate the reliability of acceptability judgments to the researchers' language backgrounds. In the Chinese context, this may refer to authors of different varieties of Mandarin Chinese (Northern Mandarin, Southern Mandarin, Taiwan Mandarin (Guoyu), etc.). To verify whether the regional background of an author may influence the acceptability judgments in published articles, we took the first step to measure this *quantitatively* by coding the background of the first author of each article we sampled. Our results show that the first author's region is not a significant factor (see Table 3). For example, sentences from journal articles whose

first authors have a Chinese mainland background do not differ in ratings from those in articles whose first authors have a non-Chinese-mainland background. These results affirm, based on our sampling, that syntacticians working on Chinese and who are speakers of different regional varieties have been reliably targeting the same grammar of Mandarin Chinese.

However, our method has some potential limitations. One is the assumption that the first author provided the grammaticality judgments for their examples. Another is the adoption of only two levels in the coding of the author backgrounds, mainland and non-mainland, due to the limited number of authors from non-mainland Mandarin-speaking communities (see Section 2.5). However, as a first step in measuring the influence of linguists' language backgrounds, we believe this operationalization can at least shed light on whether authors from the Chinese mainland produce qualitatively different examples from authors of different backgrounds. Future studies could improve on this method by including a larger group of authors from more diverse backgrounds.

5.3. Task-related factor: sentence length

As explained in Section 2.6.1, we ran a mixed-effects model that examined all factors that we consider as potential factors affecting the judgments. As shown in Table 3, the results reveal a significant interaction between sentence length and grammaticality; for grammatical sentences, the longer the sentence, the lower the acceptability rating. In other words, grammatical sentences with a longer length, as indicated by the number of characters, are more likely to be judged as "less good" or ungrammatical. However, this effect was not observed in ungrammatical sentences. This finding suggests that for ungrammatical sentences, ungrammaticality precedes parsability. In other words, parsability and grammaticality may be separate factors affecting acceptability judgments. For any given sentence, grammaticality is more likely to be binary than parsability. Parsability becomes a significant factor only for sentences that are deemed grammatical.

This finding is largely consistent with what has been reported in the literature, such as in Yao et al. (2022), who state that acceptability judgments are subject to the parsability of a sentence; complex sentences tend to add parsing difficulties for participants and are more likely to be judged as ungrammatical (Bever, 1970). In our case, this parsing difficulty is manifested through the length of the sentences and only in cases of grammatical sentences. In an experimental setting, longer sentences appear to be more cognitively demanding such that sentences of a longer length tend to be considered less natural.

5.4. Categorization of the non-replicated pairs

Twenty-seven pairs were not replicated in the LS rating task or the FC task, in at least one region. The authors of this paper together went over each pair and categorized the pairs into three groups:

Group (1): Pairs that involved structural ambiguity, inappropriate lexical item selection, or pragmatics of the example sentence. In other words, if the sentence had been constructed more carefully, if another lexical item had been chosen, or if a context had been included, the judgment may have been replicated. (N = 8)

Group (2): Pairs that were truly problematic, which may undermine the authors' theoretical claims. (N = 16)

Group (3): Pairs that should have been excluded from the stimuli (N = 3).[9]

In this section, we provide general explanations as to why these pairs failed to replicate. Further research is needed to examine each phenomenon more thoroughly to decide what factor(s) affect the acceptability judgments and what insights these results may provide to the theoretical accounts and consequent modifications thereof.

For Group (1), 8 out of the 27 pairs are categorized as involving other factors that do not necessarily undermine the authors' theories. These account for 3% of all stimuli. Two examples involve structural ambiguity. For example, the structure of one of the ungrammatical sentences is ambiguous and can be parsed as involving a structure unintended by the author. We conjecture that this ungrammatical sentence received a higher rating because of its unintended structure. Another example of ambiguity was lexical, where the monosyllabic verb *huí* has two potential meanings, that is, "to return" and "to reply," as shown in (7). The author was making the case that non-manner verbs such as *huí* cannot stand alone. For us, the unacceptability of (7-b) would only hold if *huí* was interpreted as "to return," but not "to reply." Because it is difficult to tell which meaning was interpreted among participants without a context provided for the sentence, we categorize this pair into the ambiguity group. Meanwhile, the good counterpart (7-a) also sounded odd because of the choice of the verb *pá* "to crawl." These reasons may have contributed to our participants' having judgments different from those of the linguists.

[9] These three pairs are as follows. One pair was derived from the footnote of an article where the author acknowledged she did not have a full account of the grammaticality of the pair since it was suggested by a reviewer. One pair was provided with a context by the author in the original text, whereas this context was not included in our study due to the experimental format. One pair contained a typo.

7 Pair: 38. Problem with the lexical items *pá* and *huí*. (Liu et al., 2015)

7a	Text	nǐ	rènzhende	pá.
	Gloss	you	carefully	crawl
	Trnsltn.	'You carefully crawl.'		
7b	Text	*nǐ	rènzhende	huí.
	Gloss	you	carefully	return
	Trnsltn.	'You carefully return.'		

For Group (2), 16 out of the 27 pairs are categorized as showing that the authors' theoretical claims may be problematic, since revising the structure, semantics, and pragmatics of the sentences does not seem to change the judgment (this was verified by all authors of the paper). These account for 4% of all stimuli. These include examples related to the licensing of negative polarity items (NPIs), adversity passive voice, topic construction, and focus structure, among others. We present an example involving an NPI here as an illustration.

8 Pair 254. Judgments are from the original article (same as below). In this case, the judgments are *not* replicated in our study. (Yuan, 2014)

8a	Text	zhongguó	gudài	cónglái	fángzhi	rénkouliúdòng.	
	Gloss	China	ancient	ever	prevent	population-movement	
	Trnsltn.	'Ancient China has always prevented population movement.'					
8b	Text	*zhongguó	gudài	cónglái	méiyǒu	fángzhi	rénkouliúdòng
	Gloss	China	ancient	ever	not	prevent	population-movement
	Trnsltn.	'Ancient China has never prevented population movement.'					

In (8-a), the author claimed that *prevent*, as an implicitly negative word, should be able to license the NPI *ever*, and that two licensors—*prevent* and *not*—would lead to the ungrammaticality of (8-b) because the monotonicity of the context was flipped twice and ended up in an upward-entailing environment, which should not be able to license the NPI *ever*. However, our results show the opposite in that almost all participants prefer (8-b) to (8-a). As for the theoretical source of this contrast, it remains an open question whether *prevent* is strong enough to license a strong NPI, such as *ever*, or whether the structural relationship between them leads to the unacceptability of (8-a). Our results demonstrate the importance of asking for verification from more speakers, rather than relying solely on linguists' judgments.

Note that, because each data point is associated with a specific syntactic phenomenon and theory, it is infeasible to fully discuss each of the 16 cases. We leave this for future work, which could investigate the theoretical significance of these judgments for the proposed syntactic accounts. Interested readers can find these pairs in the `osf` repository.

We also note that some cases in Group (2) have been judged differently in the literature. One of them is shown in (9), which is the passive structure in Chinese with the most common passivizer *bèi*. It has long been noted that passives in Chinese and other East Asian languages, such as Japanese and Korean, involve an indication that the resultant event is undesirable or unfortunate (e.g., Chao, 1968; Li and Thompson, 1981). As such, only verbs with negative denotations are expected to appear in passive structures. For example, compare *pīpíng* "criticize" vs.*biǎoyáng* "praise" in (9), which Liu, 2011 claims form a minimal pair.

9 Pair 25. The following contrast is not replicated. (Liu, 2011)

9a	Text	Wo	bèi	pipíng-le.
	Gloss	1sg	BEI/PASS	criticize-PERF
	Trnsltn.	'I was criticized.'		
9b	Text	?*Wo	bèi	biaoyáng-le.
	Gloss	1sg	BEI/PASS	praise-PERF
	Trnsltn.	'I was praised.'		

However, our results from both the LS rating task and the FC task show that participants from Beijing and Guangzhou consider (9-b) to be acceptable. In fact, in the LS rating task, the Beijing and Guangzhou participants gave (9-b) (6.9 out of 7) a higher rating than (9-a) (6.8 out of 7). This is in line with some other recent analyses of *bèi* passives, which

suggest that they can appear in non-negative contexts (Shao and Zhao, 2005; Xiao et al., 2006). Thus, our results suggest that the use of *bèi* passives may have shifted from being associated with negativity to neutrality because native speakers seem to agree that *bèi* passives can appear in both negative and positive contexts.

5.5. Implications for native-speakerism in research practices in syntax

The conceptualizations of ideal (native) speakers and ideal languages have up to this point greatly shaped the way we study language as a scientific enterprise. Recently, there has been a growing recognition of the need to address and challenge essentialist categories such as "monolingual" or "native speaker" in order to promote equity and inclusivity in the language sciences (Dewaele, 2018; Hackert, 2012; Bonfiglio, 2010). The essentialist categorizations of language have been shown to bias research attention toward idealized key populations, who are usually the more privileged, prototypical members of a community who have access to the so-called standard version of a language, usually located in the political center of a nation and representing the educated elites in society. Such assumptions are both harmful and inaccurate (Namboodiripad et al., 2023). Our results provide empirical evidence that challenges these essentialist ideas underlying native-speakerism. We demonstrate that Cantonese-Mandarin bilingual speakers are fully capable of providing valid acceptability judgments on Mandarin syntax. Importantly, their judgments are qualitatively identical to those of so-called ideal, monolingual Chinese speakers. These participants, who may have been considered unqualified for the task, can actually provide very important insights into the research we are undertaking. Thus, for both researchers and language users, it is time to rethink the idea of "nativeness." When researching grammaticality, researchers need to challenge assumptions about speakers' competence and should be more inclusive when recruiting participants for language studies.

6. CONCLUSION

Using stimuli randomly sampled from 10 academic journals on Chinese syntax, our study involved three experiments to probe how language background, together with other external factors such as the first author's region, article language, and sentence length, influence syntactic acceptability judgments in Mandarin Chinese.

Our results demonstrate that the judgments for Chinese sentences from journal articles are reliable overall, given the high convergence rates between the judgments from our rating experiments and those in the published articles. In addition, the acceptability judgments made by Mandarin-Cantonese bilinguals do not differ significantly from those made by monolingual Mandarin speakers, despite the fact that the monolingual speakers tend to make crisper decisions on good vs. bad sentences in their judgments. This indicates that the grammar in a bilingual mind may be different from the grammar in a monolingual mind in terms of quantity (i.e., gradient) rather than quality (i.e., categorical), as their judgments largely converged. Therefore, we urge the community to rethink the concept of an ideal native speaker and advocate more inclusive recruitment criteria in language studies.

That said, our conclusions should be interpreted with caution. As a reviewer pointed out, the promotion of Putonghua (i.e., Standard Mandarin) nationwide over the last 70 years has made it a common and universal language in China. Consequently, Putonghua has largely replaced regional dialects as the first language in many areas. In some places, even if it is not the primary language, it serves as the medium of instruction in all schools. Children growing up in these environments are often bilingual and share similar language intuitions with native speakers. The participants from Guangzhou in this study are likely part of this group. Furthermore, Guangzhou is a large, developed city with residents from all over China, where Putonghua is widely spoken. Therefore, our conclusions may not hold for participants from remote, dialect-speaking areas, where Putonghua is truly a second language. Future research should explore whether our findings are generalizable to acceptability judgments in different contexts of bilingualism.

7. FUNDING

This work is funded by the Humanities and Social Sciences Grant from the Chinese Ministry of Education (No. 22YJC740020) awarded to Hai Hu, as well as the Shanghai Pujiang Program Grant (No. 22PJC063) awarded to Hai Hu. The research was also sponsored by the research fund of Chien-Jer Charles Lin from Indiana University Bloomington.

Declaration of Competing Interest

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

ACKNOWLEDGMENTS

We thank Peng Zhang, Liceng Liu, Yaxin Liu, Yue Xu, Chun Zheng, Xiaojing Zhao, Yushu Wang, Zihan Zhao for their help in data collection and annotation. We are also grateful to Zhong Chen, Jackie Lai, Jon Sprouse for discussions on early drafts of the paper, and Gwendolyn Hildebrandt for proofreading the paper. This paper is dedicated to Jiahui Huang.

APPENDIX A

A.1. Selected minimal pairs from the stimuli

A.1.1. Selected minimal pairs from Experiment 1

We list ten sentences (five minimal pairs) randomly sampled from the stimuli in Experiment 1, which is a LS rating task.[10] For each sentence, we provide the mean ratings from the participants in Beijing (BJ) and Guangzhou (GZ).

10		Pair ID: 32 (Hwang and Tai, 2014); LS ratings: 4.9 vs. 2.7 (BJ) and 5.5 vs. 3.6 (GZ)		
10a	Text	ta	da-zhe	qiú.
	Gloss	3sg	hit-DUR	ball
	Trnsltn.	'S/he is playing basketball.'		
10b	Text	*ta	chéngshí-zhe.	
	Gloss	3sg	honest-DUR	
	Trnsltn.	'*S/he is honesting.'		

11		Pair ID: 60 (Wei, 2011); LS ratings: 4.3 vs. 4.0 (BJ) and 5.6 vs. 4.3 (GZ)								
11a	Text	zhangsan	kàndào	mourén,	dàn	wo	bù	zhidào	shì	shuí.
	Gloss	Zhangsan	see	someone	but	I	NEG	know	be	who
	Trnsltn.	'Zhangsan saw someone, but I don't know who that was.'								
11b	Text	*zhangsan	kàndào	mourén,	dàn	wo	bù	zhi	shuí.	
	Gloss	Zhangsan	see	someone	but	I	NEG	know	who	
	Trnsltn.	'Zhangsan saw someone, but I don't know who that was.'								

12		Pair ID: 68 (Yang, 2011); LS ratings: 6.2 vs. 1.9 (BJ) and 6.2 vs. 2.3 (GZ)									
12a	Text	tā fàng-zài-le	zài zhuōzi	yī shàng	gè	xiǎoshí	nèi	bǎ	suǒyǒu	de	shū
	Gloss	3sg place-at-PERF	at table	one on	CL	hour	in	BA	all	DE	book
	Trnsltn.	'S/he placed all the books on the table in an hour.'									
12b	Text	*tā	zài	yī	gè	xiǎoshí	nèi	zhàn-le.			
	Gloss	3sg	at	one	CL	hour	within	stand-PERF			
	Trnsltn.	'S/he stood in an hour.'									

[10] The abbreviations used in the gloss of the following examples include: CL for classifier; DE for modifier-modified marker; DUR for durative marker; PERF for perfective marker; PROG for progressive marker; SFP for sentence final particle; BA for the BA construction; BEI for the BEI construction; GEI for introducing an external force to the verb.

13		Pair ID: 130 (Huang, 2012); LS ratings: 4.8 vs. 1.8 (BJ) and 5.1 vs. 1.5 (GZ)							
13a	Text	yǒu	yi	gè	rén	xǐhuān	lisì	ma?	
	Gloss	have	one	CL	person	like	Lisi	SFP	
	Trnsltn.	'Is there anyone that likes Lisi?'							
13b	Text	*yǒu	yi	gè	rén	xǐ	bù	xǐhuān	lisì?
	Gloss	have	one	CL	person	like	not	like	Lisi
	Trnsltn.	'Is there anyone that likes Lisi or not?'							

14		Pair ID: 291 (Shen and Rint, 2010); LS ratings: 6.6 vs. 2.2 (BJ) and 6.1 vs. 2.8 (GZ)						
14a	Text	nà	fú	huà	bèi	tā	gei	mài-le.
	Gloss	that	CL	picture	BEI	3sg	GEI	sell-PERF
	Trnsltn.	'That picture was sold (by someone).'						
14b	Text	*nà	fú	huà	bèi	tā	bèi	mài-le.
	Gloss	that	CL	picture	BEI	3sg	BEI	sell-PERF
	Trnsltn.	'That picture was sold.'						

A.1.2. Selected minimal pairs from Experiment 2 (test group)

We list five randomly sampled minimal pairs from the stimuli in Experiment 2, which is an FC task. We also present the choices of the participants for these pairs in Fig. 9.

15		Pair ID: 138 (Xie, 2015)							
15a	Text	zhāngsān	zuótian	hái	zuò	de	wán	nàxiē	zuòyè.
	Gloss	Zhangsan	yesterday	still	do	ability.modal	finish	those	homework
	Trnsltn.	'Zhangsan still had the ability to finish that homework yesterday.'							
15b	Text	*zhāngsān	zuótian	zuò	de	wán	nàxiē	zuòyè.	
	Gloss	Zhangsan	yesterday	do	DE	finish	those	homework	
	Trnsltn.	'Zhangsan had the ability to finish that homework yesterday.'							

16		Pair ID: 150 (Yang, 2017)						
16a	Text	zhè	gè	dìfāng	zuótiān	huòxǔ	hái	ānquán.
	Gloss	this	CL	place	yesterday	maybe	still	safe
	Trnsltn.	'This place may be still safe yesterday.'						
16b	Text	*zhè	gè	dìfāng	zuótiān	hái	huòxǔ	ānquán.
	Gloss	this	CL	place	yesterday	still	maybe	safe
	Trnsltn.	'This place may be still safe yesterday.'						

17		Pair ID: 272 (Zhou and Chen, 2013)								
17a	Text	yī	gè	rìběn	jūnguān	bǐbǐhuàhuà	de	jiǎng-zhe	rìběn	huà.
	Gloss	one	CL	japan	officer	gesture	DE	speak-DUR	japan	word
	Trnsltn.	'A Japanese officer was speaking Japanese while gesturing.'								
17b	Text	*yī	gè	rìběn	jūnguān	bǐbǐhuàhuà	de	zài	jiǎng	shénme?
	Gloss	one	CL	japan	officer	gesture	DE	PROG	speak	SFP
	Trnsltn.	'What was a Japanese officer speaking while gesturing?"								

18		Pair ID: 23 (Tsao, 2010)									
18a	Text	zhè	shì	nǐ	sīzì	duì	nà	jiàn	shì	de	pīpíng.
	Gloss	this	is	2sg	private	to	that	CL	matter	DE	criticism
	Trnsltn.	'This is your private criticism of that matter.'									
18b	Text	*zhè	shì	nǐ	duì	nà	jiàn	shì	de	sizì	pīpíng.
	Gloss	this	is	2sg	to	that	CL	matter	DE	private	criticism
	Trnsltn.	'This is your private criticism of that matter.'									

19		Pair ID: 286 (Li, 2011)										
19a	Text	nǐ	cháo-zhe	dírén	hōng	yī	pào,	tāmen	jiù	xià	pǎo-le.	
	Gloss	2sg	at-DUR	enemy	bomb	one	cannon	3pl	then	scare	run-PERF	
	Trnsltn.	'You fire a cannon at the enemy and then they will get scared away.'										
19b	Text	*nǐ	cháo-zhe	dírén	hōng	yī	huí	pào,	tāmen	jiù	xiàpǎo	le.
	Gloss	2sg	at-DUR	enemy	bomb	one	time	cannon	3pl	then	scare	run-PERF
	Trnsltn.	'You fire a cannon once at the enemy and then they will get scared away.'										

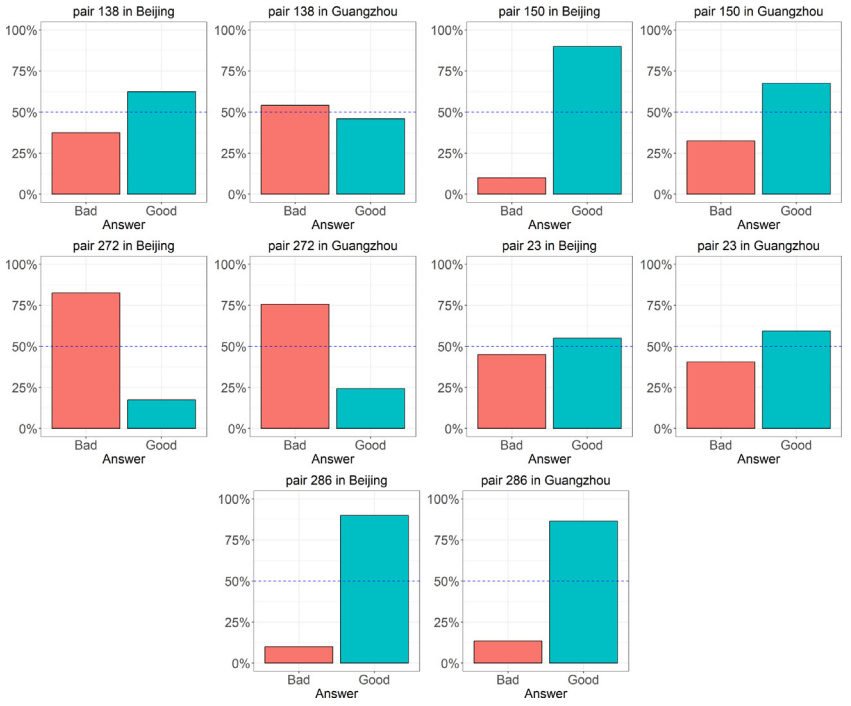

Fig. 9. Choices of the participants from the two regions for the five pairs from Experiment 2.

A.1.3. Selected minimal pairs from Experiment 3 (in notRepBJonly and notRepGZonly groups)

We list five randomly sampled minimal pairs from the stimuli in Experiment 3, which is also an FC task. We also present the choices of the participants in Fig. 10.

20		Pair ID: 301 (Wan, 2011)						
20a	Text	nǐ	zhème	yī	shuō,	wǒ	zhīdào-le.	
	Gloss	2sg	such	one	speak	I	know-PERF	
	Trnsltn.	'Now that you said like that, I got it.'						
20b	Text	*nǐ	zhème	yī	shuō,	wǒ	zhīdào	de.
	Gloss	2sg	such	one	speak	I	know	DE
	Trnsltn.	'Now that you said like that, I got it.'						

21		Pair ID: 1 (Fan and Li, 2019)			
21a	Text	kēngr	tāmen	wā	qiǎn-le.
	Gloss	hole	3pl	dig	shallow-PERF
	Trnsltn.	'The hole, they have dug it too shallow.'			
21b	Text	*tāmen	wā	kēngr	qiǎn-le.
	Gloss	3pl	dig	hole	shallow-PERF
	Trnsltn.	'The hole, they have dug it too shallow.'			

22		Pair ID: 287 (Li, 2011)											
22a	Text	nǐ	xiàng	nà	gè	wān	de	gùnzi	qiāo	yi	tiěqián,	jiù	zhí-le.
	Gloss	2sg	at	that	CL	curve	DE	stick	hit	one	iron.plier	then	straight-PERF
	Trnsltn.	'You hit the curved stick with a pair of iron pliers and it will straighten.'											
22b	Text	*nǐ	xiàng	nà	gè	wān	de	gùnzi	qiāo	yi	yìngwù,	jiù	zhí-le.
	Gloss	2sg	at	that	CL	curve	DE	stick	hit	one	hard.thing	then	straight-PERF
	Trnsltn.	'You hit the curved stick with a hard thing and it will straighten.'											

23		Pair ID: 233 (Huang, 2012)									
23a	Text	zhè gāo dù	zhāng dà	zhuōzi yīxiē.	de	chángdù	bǐ	nà	gè	shujià	de
	Gloss	this height	CL big	table some	DE	length	compare	that	CL	bookshelf	DE
	Trnsltn.	'The length of this table is slightly longer than the height of that bookshelf.'									
23b	Text	*zhè	zhāng	zhuōzi	bǐ	nà	gè	shujià	gāo	gāo	yīxiē.
	Gloss	this	CL	table	compare	that	CL	bookshelf	very	big	some
	Trnsltn.	'This table is a bit higher than that bookshelf.'									

24		Pair ID: 96 (Cheng and Sybesma, 1999)						
24a	Text	tā	xie-guò	yì	běn	shū	hěn	yǒu-yìsī.
	Gloss	3sg	write-EXP	one	CL	book	very	interesting
	Trnsltn.	'S/he once wrote a book which was very interesting.'						
24b	Text	*tā	xie-guò	běn	shū	hěn	yǒu-yìsī.	
	Gloss	3sg	write-EXP	CL	book	very	interesting	
	Trnsltn.	'S/he once wrote a book which was very interesting.'						

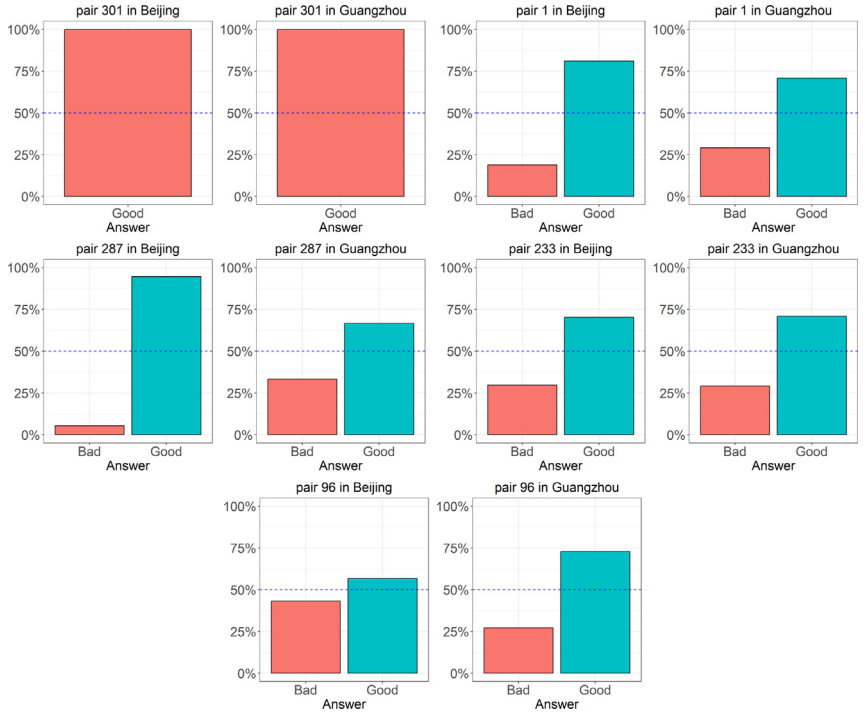

Fig. 10. Choices of the participants from the two regions for the five pairs from Experiment 3.

A.2. Post-hoc comparisons for Experiment 3

Post-hoc comparisons for Experiment 3 are presented in Table 9.

Table 9: Post-hoc pairwise comparisons in Experiment 3.

Condition	Contrast	Estimate	Std.Error	z-value	Pr(>|z|)
Control	Beijing - Guangzhou	0.88	0.33	2.65	<0.01
notRepBJonly	Beijing - Guangzhou	−0.23	0.18	−1.30	0.20
notRepGZonly	Beijing - Guangzhou	0.82	0.22	3.74	<0.001
Beijing	Control - notRepBJonly	2.61	0.54	4.85	<0.001
Beijing	Control - notRepGZonly	1.95	0.53	3.68	0.001
Beijing	notRepBJonly - notRepGZonly	−0.66	0.44	−1.49	0.30
Guangzhou	Control - notRepBJonly	1.50	0.50	3.02	0.01
Guangzhou	Control - notRepGZonly	1.89	0.48	3.99	<0.001
Guangzhou	notRepBJonly - notRepGZonly	0.39	0.42	0.93	0.62

References

Barbiers, S., Bennis, H., 2007. The syntactic atlas of the Dutch dialects. Nordlyd 34 (1).
Bates, D., Maechler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67, 1–48.
Bever, T.G., 1970. The cognitive basis for linguistic structures. In: Hayes, J.R. (Ed.), Cognition and the development of language. John Wiley and Sons Inc, New York.

Bonfiglio, T.P., 2010. Mother tongues and nations: the invention of the native speaker. De Gruyter Mouton.
Chao, Y.-R., 1968. A grammar of spoken Chinese. University of California Press, Berkeley.
Chaves, R. P. and Dery, J. E. (2014). Which subject islands will the acceptability of improve with repeated exposure? In Chaves, R. P. and Dery, J. E., editors, Proceedings of the 31st West Coast Conference on Formal Linguistics, pages 96–106, Somerville, MA: Cascadilla Proceedings Project.
Chaves, R. P. and Jeruen E., D. (2018). Frequency effects in subject islands. Journal of Linguistics, 55(3):475–521.
Chen, Z., Xu, Y., Xie, Z., 2020. Assessing introspective linguistic judgments quantitatively: the case of *The Syntax of Chinese*. J. East Asian Linguis. 29 (3), 311–336.
Cheng, L.L.-S., Sybesma, R., 1999. Bare and not-so-bare nouns and the structure of NP. Linguistic Inquiry 30 (4), 509–542.
Dewaele, J.-M., 2018. Why the dichotomy "L1 versus LX user" is better than "native versus non-native speaker". Appl. Linguistics 39 (2), 236–240.
Edelstein, E., 2014. This syntax needs studied. In: Zanuttini, R., Horn, L.R. (Eds.), Micro-syntactic variation in North American English, Oxford Studies in Comparative Syntax. Oxford University Press, Oxford, pp. 242–268.
Fan, Y., Li, M., 2019. The obligatory fronting of the undergoer argument in the Mandarin excessive serial verb construction. Concentric 45 (2), 167–191.
Francom, J., 2009. Experimental Syntax: Exploring the Effect of Repeated Exposure to Anoma-lous Syntactic Structure—Evidence from Rating and Reading Tasks PhD thesis. University of Arizona.
Hackert, S., 2012. The emergence of the English native speaker: A chapter in nineteenth-century linguistic thought,, volume 4. Walter de Gruyter.
Hofmeister, P., Casasanto, L.S., Sag, I.A., 2012. How do individual cognitive differences relate to acceptability judgments? a reply to Sprouse, Wagers, and Phillips. Language 88 (2), 390–400.
Hofmeister, P., Norcliffe, E., 2013. Does resumption facilitate sentence comprehension? In: Hofmeister, P., Norcliffe, E. (Eds.), The Core and the Periphery: Data-Driven Perspectives on Syntax Inspired. CSLI Publications, Stanford, CA, pp. 225–246.
Huang, R.-H.R., 2012. On two types of existential subjects in Chinese A-not-A questions. Lang. Linguist. 13 (6), 1171.
Hwang, H.-H., Tai, J.H., 2014. Temporal sequence structure and the aspect marker-zhe in Chinese. J. Chin. Linguist. 42 (1), 39–54.
Kutlu, E., Chiu, S., McMurray, B., 2022. Moving away from deficiency models: Gradiency in bilingual speech categorization. Front. Psychol. 13.
Li, C.N., Thompson, S.A., 1981. Mandarin Chinese: a functional reference grammar. University of California Press, Berkeley.
Li, F.-K., 1973. Languages and dialects of China. Journal of Chinese Linguistics, 1–13.
Li, X., 2011. Can verb classifiers be borrowed from nouns?. Zhongguo Yuwen (Chinese Philology) 343 (4), 87–114.
Liao, W.-W.R., Wang, Y.I., 2011. Multiple-classifier constructions and nominal expressions in Chinese. J. East Asian Linguis. 20, 145–168.
Lin, C.-J.C., 2018. Subject prominence and processing dependencies in prenominal relative clauses: The comprehension of possessive relative clauses and adjunct relative clauses in Mandarin Chinese. Language 94 (4), 758–797.
Linzen, T., Oseki, Y., 2018. The reliability of acceptability judgments across languages. Glossa: a J. General Linguist. 3 (1), 1–25.
Liu, H., 2011. Expletive negation in Mandarin cha-dian-mei 'miss-bit-not' + V structure. J. Chin. Linguist. 39 (1), 123–148.
Liu, M., Tsai, H., Hu, C., Chou, S., 2015. The proto-motion event schema: Integrating lexical semantics and morphological sequencing. J. Chin. Linguist. 43 (2), 503–547.
Mahowald, K., Graff, P., Hartman, J., Gibson, E., 2016. SNAP judgments: a small N acceptability paradigm (SNAP) for linguistic acceptability judgments. Language 92 (3), 619–635.
Murray, T.E., Frazer, T.C., Simon, B.L., 1996. Need + past participle in American English. American Speech 71 (3), 255–271.
Myers, J., 2009. Syntactic judgment experiments. Lang. Linguist. Compass 3 (1), 406–423.
Namboodiripad, S., Kutlu, E., Babel, A., Babel, M., Baese-Berk, M., Bassuk, P. B., Block, A., Carlson, M., Cheng, A., Combiths, P., et al. (2023). Essentialist characterizations of language are an obstacle to accuracy, progress, and justice in science.
Phillips, C., 2013. Some arguments and nonarguments for reductionist accounts of syntactic phenomena. Lang. Cognit. Process. 28 (1–2), 156–187.
Poletto, C., Benincà, P., 2007. The ASIS enterprise: a view on the construction of a syntactic atlas for the Northern Italian dialects. Nordlyd 34 (1).
R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Schütze, C.T., 2016. The empirical base of linguistics: Grammaticality judgments and linguistic methodology. Language Science Press.
Shao, J., Zhao, C., 2005. Cognitive interpretation of *ba*-construction and *bei*-construction. Chin. Lang. Learn. 4, 11–18.
Shen, Y., Rint, S., 2010. The derivational relation between the syntactic marker *gei* and several verbal constructions. Zhongguo Yuwen (Chinese Philology) 336 (3).
Snyder, W., 2000. An experimental investigation of syntactic satiation effects. Linguist. Inquiry 31 (3), 575–582.
Sprouse, J. (2018). Acceptability judgments and grammaticality, prospects and challenges. In Hornstein, N., Yang, C., and Patel-Grosz, P., editors, Syntactic Structures after 60 Years: The Impact of the Chomskyan Revolution in Linguistics, volume 60, pages 195–224.
Sprouse, J., Almeida, D., 2012. Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. J. Linguist., 609–652.

Sprouse, J., Almeida, D., 2017. Design sensitivity and statistical power in acceptability judgment experiments. Glossa: A J. General Linguist. 2 (1), 1–32.
Sprouse, J., Schütze, C.T., Almeida, D., 2013. A comparison of informal and formal acceptability judgments using a random sample from *Linguistic Inquiry* 2001–2010. Lingua 134, 219–248.
Sprouse, J., Wagers, M., Phillips, C., 2012. A test of the relation between working-memory capacity and syntactic island effects. Language 88 (1), 82–123.
Tang, C., van Heuven, V.J., 2007. Mutual intelligibility and similarity of Chinese dialects: predicting judgments from objective measures. Linguist. Netherlands 24 (1), 223–234.
Tang, C., Van Heuven, V.J., 2009. Mutual intelligibility of Chinese dialects experimentally tested. Lingua 119 (5), 709–732.
Tsao, F.-F., 2010. Complement and adjunct distribution and the two-place nominals in Chinese NPs. J. Chin. Linguist. 38 (1), 87–114.
Wan, Q., 2011. De in state-of-affairs sentences. Zhongguo Yuwen (Chinese Philol.) 352 (1), 87–114.
Wang, C., Liu, W., 2014. On Chinese perfective marker *le*: a syntactic analysis based on minimalist program. Yuyan Kexue (Lang. Sci.) 13 (4), 355–368.
Wei, T.-C., 2011. Island repair effects of the left branch condition in Mandarin Chinese. J. East Asian Linguis. 20, 255–289.
Weskott, T., Fanselow, G., 2011. On the informativity of different measures of linguistic acceptability. Language, 249–273.
Wickham, H., 2011. ggplot2. Wiley Interdisciplinary Reviews: Computational Statistics 3 (2), 180–185.
Xiao, Z.R., McEnery, A., Qian, Y., 2006. Passive constructions in English and Chinese: a corpus-based contrastive study. Lang. Contrast 6, 109–149.
Xie, Z., 2015. Non-root modals for the past and temporal shifting in Mandarin Chinese. Lingua Sinica 1, 1–22.
Yang, C.-Y.H., 2017. On the syntax-semantics interface of focus particles: the additive particle *hai* in Mandarin Chinese. Lingua Sinica 3, 1–33.
Yang, S., 2011. The parameter of temporal endpoint and the basic function of *le*. J. East Asian Linguis. 20, 383–415.
Yao, Y., Xie, Z., Lin, C.-J. C., Huang, C.-R., 2022. Grammatical acceptability in Mandarin Chinese. In Huang, C.-R., Lin, Y.-H., and Chen, I.-H., editors, Cambridge Handbook of Chinese Linguistics, pages 669–706.
Yuan, Y., 2014. Factivity and polarity licensing function of implicitly negative verbs. Yuyan Kexue (Lang. Sci.) 13 (6), 575–586.
Zanuttini, R., Wood, J., Zentz, J., Horn, L., 2018. The Yale grammatical diversity project: morphosyntactic variation in North American English. Linguistics Vanguard 4 (1).
Zhou, S., Chen, Z., 2013. A quantitative study on the licensing conditions of indefinite subjects marked by yi + quantifier + noun. Yuyan Kexue (Lang. Sci.) 12 (4), 371–382.
Zhou, Y., Jiang, H., 2014. The nature and licensing conditions of three elliptical structures in subsequent sentences in Mandarin. Yuyan Kexue (Lang. Sci.) 6, 601–614.