To appear in *Journal of East Asian Linguistics*

May 20, 2020

Email: zxcgsl@rit.edu

Assessing introspective linguistic judgments quantitatively: The case of *The Syntax of Chinese*

Zhong Chen
Rochester Institute of Technology

Yuhang Xu
University of Rochester

Zhiguo Xie
The Ohio State University

Abstract

The informal judgments of the well-formedness of phrases and sentences have long been used as the primary data source for syntacticians. In recent years, the reliability of data based on linguists' introspective intuitions is increasingly subject to scrutiny. Although a number of studies were able to replicate a vast majority of English judgments published in a textbook and in peer-reviewed journal articles, the status of data in many non-English languages has yet to be experimentally examined. In this work, we employed formal quantitative methods to evaluate the reliability of judgments in the widely used textbook, *The Syntax of Chinese* (Huang, Li, & Li, 2009). We first assessed example sentences based on the acceptability ratings from 148 native Mandarin Chinese speakers. Using a target forced-choice task, we further explored the potentially problematic sentence pairs. Results of the two experiments suggest an eminently successful replication of judgments in the book: out of the 557 data samples tested, only five sentence pairs require further investigation. This large-scale study represents the first attempt to replicate the judgments in a non-English syntax textbook, in hopes to bridge the gap between the informal data-collection in Chinese linguistic research and the protocols of experimental cognitive science.

Keywords: Acceptability judgments, Reliability, Experimental syntax, Chinese

1 Introduction

Over the past 50 years, the empirical base of forming syntactic theories has largely relied on acceptability judgments (Chomsky, 1965; Schütze, 1996). Sentences are constructed, compared, and discussed with respect to whether they are grammatically acceptable. The data collection process, however, is often introspective and informal, which reflects the judgments of only the researcher(s) and occasionally feedback from colleagues and a small number of "naïve" speakers. Researchers have long been asking questions about the grammaticality-acceptablility relationship and the reliance on individual syntacticians' intuition as opposed to consulting native speakers of the target language (Birdsong, 1989; Edelman & Christiansen, 2003; Labov, 1978; Langendoen, Kalish-Landon, & Dore, 1973; Levelt, van Gent, Haans, & Meijers, 1977; Newmeyer, 1983; Schütze, 1996, among others). In recent years, the reliability of judgments and the syntactic theories which they support are increasingly subject to scrutiny (e.g. Edelman & Christiansen, 2003; Gibson & Fedorenko, 2010; Gibson, Piantadosi, & Fedorenko, 2013). The critics have largely targeted the introspective judgments on the basis of not following prevalent methodological protocols adopted in many other fields (e.g. experimental psychology), and a lack of applying statistical techniques to the analyses.

A number of formal approaches to test sentence acceptability have been proposed to replicate judgments taken from the syntax literature, most of which employ internet-based tools that are freely available. Crowdsourcing platforms such as Amazon Mechanical Turk further allow us to collect high quality judgment data efficiently as an alternative to laboratory-based studies (Munro et al., 2010; Sprouse, 2011).[1] Native speakers are recruited to participate in large-scale online experiments of rating sentences on a Likert scale, using magnitude estimation, or making a forced-choice (Mahowald et al., 2016; Sprouse & Almeida, 2012; Sprouse, Schütze, & Almeida, 2013, among others).[2] Two representative studies by Sprouse and colleagues replicated the vast majority of English judgments in the widely used textbook *Core Syntax* (Adger, 2003) and from articles published in the journal *Linguistic Inquiry* between 2001 and 2010. The results suggest a strong convergence between linguists' judgments in the syntax literature and those of non-linguist native speakers of English (Sprouse & Almeida, 2012; Sprouse et al., 2013). Although there are still concerns about whether the reported error rate of replication is low enough (Gibson et al., 2013), those results lend support to the idea that introspective judgments by individual syntacticians and their small number of consultants, if any, remain to be reliable and valuable (den Dikken, Bernstein, Tortora, & Zanuttini, 2007; Phillips, 2009; Phillips & Lasnik, 2003). Proponents of formal methods point out that assessing informal judgments quantitatively would allow us to compare the size of effect in different constructions and evaluate the consistency of judgments across different studies and languages (Cowart, 1997; Featherston, 2005; Keller, 2000; Mahowald et al., 2016; Phillips, 2009). Sprouse and colleagues,

[1] It is also possible to obtain informative quantitative data with fewer participants using software tools like the one illustrated in Myers (2009a) or using the Bayesian-framework paradigm proposed by Mahowald, Graff, Hartman, and Gibson (2016).

[2] Sprouse and Almeida (2017) were the first to compare the statistical power of various judgment tasks. A follow-up work by Langsford, Perfors, Hendrickson, Kennedy, and Navarro (2018) estimated how much of the variability within each task is due to psychometric properties, including participant-level individual differences, sample size, response styles, and item effects.

in particular, have reported case studies for phenomena like the island effects which are directly related to sentence acceptability (Sprouse, 2011; Sprouse, Wagers, & Phillips, 2012).

While English accounts for a large proportion of data sources in linguistic research, other languages have undoubtedly played a key role in achieving the principal goals of generative syntax, i.e., "universalist", "particularist", and "typological" (Newmeyer, 2013). A large number of works rely directly on crosslinguistic contrast and comparison. For example, about half of the syntax-focused articles published in *Linguistic Inquiry* between 2001 and 2010 were predominantly[3] about phenomena in languages other than English (Sprouse et al., 2013, p. 222). On the contrary, studies that aim to replicate judgments in non-English syntax literature are relatively limited so far. Song, Choe, and Oh (2014) tested sentences excerpted from two volumes of a Korean linguistics journal in a five-point Likert scale task. They reported a convergence rate of 84.75% on the pairwise sentences, as compared to 95% for the English journal data reported in Sprouse et al. (2013). Linzen and Oseki (2018), on the other hand, focused on selective judgments that they believed to be questionable in journal articles about Hebrew and Japanese syntax. Although approximately half of the contrasts did not replicate in the seven-point Likert scale experiment, the high error rate in their study was not directly comparable to the results of Sprouse et al. (2013) who instead tested random samples. In Linzen and Oseki's own words, their "results do not suggest that there is a 'replicability crisis' in Hebrew or Japanese linguistics". This was mostly because the experimental materials were deliberately chosen without including dozens of non-controversial contrasts for each questionable judgment.

In this work, we expand the discussion of judgment reliability by providing experimental data on the example sentences published in a recent and frequently used Chinese syntax textbook, *The Syntax of Chinese* (by C.-T. James Huang, Y.-H. Audrey Li, and Yafei Li, 2009, Cambridge University Press, henceforth, HLL). Chinese syntactic research has contributed greatly to the development of modern syntactic theories since the 1960s and even more so since the mid-1980s. It has roughly shared the same methodology as the rest of the syntax field and largely relied on theoretical argumentation built upon introspective judgments. It was not until recently that the experimental syntactic research emerged to test the validity of empirical observations and theoretical generalizations. Most of those studies, however, focused on one language phenomenon (e.g. Zhou and Gao (2009) on scope rigidity; Lu, Thompson, and Yoshida (to appear) on wh-in-situ and island effects). Formal experimental assessments at a more global level are still lacking. Against this background, in our study we test the language data points which summarize key results on a variety of Chinese topics. The HLL book develops argumentation by carefully analyzing its examples under the generative syntax framework. It is one of the most commonly adopted textbooks in Chinese linguistics. As we discuss below in Section 2, instead of randomly sampling (or selectively choosing) from a pool of contrasts, the present research aims to assess the judgments on all *testable* example sentences in the book, in a way parallel to Sprouse and Almeida (2012). The test items include sentence contrasts illustrating uncontroversial facts about the Chinese grammar as well as those that are instrumental in forming theoretical arguments. Previous experiments on languages other than English have chosen to assess acceptability judgments from journal articles rather than from text-

[3] Sprouse et al. (2013) defined "predominantly" as more than 80% of the data points in an article.

books because "many of the judgments in the (textbook) literature are self-evident" and "are not the judgments that critics take issue with" (Linzen & Oseki, 2018, p. 2). We believe, however, as the first step, testing examples from a Chinese textbook allows us to compare our results with similar works, such as Sprouse and Almeida (2012). It also establishes a baseline for future evaluations of judgments about different topics in Chinese syntax when other sources of materials are used. Importantly, we hope that this work adds to the discussion about published judgments in English versus other languages, as it seems "premature" to draw any conclusions given the current research status (Schütze, 2020).

The rest of this paper is organized as follows. Section 2 describes an acceptability rating experiment which assessed Chinese native speakers' judgments of the sentence examples in the HLL book. Section 3 reports a forced-choice task in which we further evaluated sentence contrasts that raised a flag in the rating experiment. In Section 4, we discuss the interpretation of these results while relating our work to relevant previous studies. Section 5 concludes the paper.

2 Experiment 1: Acceptability rating

2.1 Materials

In this replication study of the HLL book, the material selection procedure started with the collection of 1,186 sentence examples in modern Mandarin Chinese by excluding others from dialects, subdialects, classical Chinese, or languages other than Chinese.[4] Using the syntactic acceptability notations in the book, we grouped sentences into three categories. The acceptable "good" sentences were unmarked, whereas the "bad" ones received either an unacceptable asterisk "*" or a highly questionable "??" symbol. The third category fell in between. It included sentences that were not fully acceptable and marked by a "?" symbol, indicating their "questionable" status.[5]

By going through each of examples, we identified the items to test in Experiment 1. We found that sentences in the book were compared in many different ways and decided to focus on the 158 PAIR contrasts, each of which contained a "good" sentence and its "bad" counterpart. They demonstrated the most typical and direct pairwise comparisons in terms of syntactic acceptability, and such pairs were the main focus of discussion in previous replication studies. We also included 42 TRIPLE contrasts in which three sentences were compared. Lastly, we tested 115 OTHER items, including 39 pairs which were not the "good versus bad" type, 5 quadruple contrasts,[6] as well as 15 independent examples to balance the number of "good" and "bad" sentences in the experiment. Three examples of the test items are shown below in (1)-(3):

(1) Example of a PAIR contrast (HLL, p. 172)

[4] While the HLL book was originally published in English, the experimental stimuli were presented in scripts from the book's simplified Chinese edition (Huang, Li, & Li, 2013) where spaces segmenting two adjacent words were removed. For all examples in this paper, the page numbers that we refer to are from the book's English edition.

[5] The experimental materials (and the excluded sentences), data, and code for this manuscript are available at `https://osf.io/374h6/`

[6] These quadruples, e.g. a group of four "bad" sentences, can not be divided into two PAIR contrasts, nor can they be analyzed using two-way ANOVAs.

a. *ta bei women kandao-le.*
he BEI us see-LE
'He was seen by us.'

b. **women ba ta kandao-le.*
we BA him see-LE
'We saw him.'

(2) Example of a TRIPLE contrast (HLL, p. 82)

a. *ta di-gei gege yi-hu jiu.*
he pass-give brother one-CL wine
'He passed his brother a jug of wine.'

b. *?ta di-gei gege yi-hu jiu, jiejie yi-pan cai.*
he pass-give brother one-CL wine sister one-CL dish
'He passed his brother a jug of wine and his sister a dish.'

c. **ta di-gei de shi gege yi-hu jiu.*
he pass-give DE be brother one-CL wine
*'What he passed was his brother a jug of wine.'

(3) Example of an OTHER - independent sentence (HLL, p. 193)

**wo ba zhe-ping-jiu he-zui-le.*
I BA this-CL-wine drink-drunk-LE

*'I have drunk the wine drunk.'

The syntactic acceptability of some types of sentences is determined by the correctness of co-reference, such as the one below (HLL, p. 330), which linguists use indices to explain.

(4) *Zhangsan*$_i$ *zhidao Lisi*$_j$ *lao piping taziji*$_{*i/j/*k}$.
Zhangsan know Lisi incessantly criticize himself
'Zhangsan knows that Lisi criticizes himself all the time.'

Following previous studies, we chose not to test example sentences with co-referential indices, including all data points from Chapter 9 "Anaphora" where most, if not all, examples are in this class. This was because the acceptability judgment experiments, such as the seven-point Likert scale rating task we used, would not be able to straightforwardly probe whether or not a reader successfully identifies a co-referential dependency. We also excluded sentences such as the following (HLL, p. 159) whose meaning needs to be contextualized, because items in our experiments were presented without any preceding context and there was no special instruction of the participants.

(5) *wo bei ta zheme yi zuo, (wo) jiu shenme dou kan-bu-jian le.*
I BEI he thus one sit I then everything all can-not-see LE
'As soon as I had him sitting this way [on me], I couldn't see anything at all.'
Context: said of a concert, when someone tall sits in front of me and blocks my view.

In the HLL book, parentheses are sometimes introduced when discussing whether the expression contained within them is optional. In our study, we treated any example with notational parentheses as two test items in which the expression within the parentheses was omitted in one of the items. For instance, the example below (HLL, p. 227) was presented to participants as a Pair contrast in which *yuanyin* "reason" was omitted in the unacceptable condition.

(6) *ni yinggai ba ta weishenme bu lai de *(yuanyin) gaosu women.*
you should BA he why not come DE reason tell us
'You should tell us the (reason) why he didn't come.'

In addition, multiple variants of the same sentence, e.g. acceptable versus unacceptable, were sometimes presented under different but normally continuous example numbers in the book. We attempted to re-group and assess those variants together within the same contrast in our experiments. For consistency, we only included the identifiers coded for our test items in figures of this paper. The mapping between item identifiers and their corresponding example numbers in the HLL book is available in the supplementary data.

Table 1
The number of test items selected from each chapter of the HLL book.

Chapter	Good	Questionable	Bad
	unmarked	“?"	“??" or “∗"
1	17	2	14
2	15	5	11
3	24	7	20
4	22	0	20
5	43	1	48
6	54	2	61
7	41	0	45
8	44	2	59
9	0	0	0
Total	260	19	278

Table 1 provides a summary of the examples selected from each chapter of the HLL book. These 557 sentences were the only items that we tested in Experiment 1. Our test items stand in contrast with those in Sprouse and Almeida (2012), who constructed 7 additional items of the same pattern for each target sentence and analyzed those 8 items as a whole. Their choice of doing so has allowed them to speak directly to the criticism of generalizability by Gibson and Fedorenko (2010), who argued that the acceptability difference between two sentences, if observed, could be a result of specific lexical bias, rather than the properties of syntactic construction. Accordingly, the judgment pattern, and ultimately the syntactic theory built on that, would have to hold for the variance in lexical properties. Formal experiments with multiple stimuli for each target construction allow linguists to test the robustness of a hypothesis. Based on multiple sets of items, Sprouse et al.'s successful replication studies lend strong support to the exemplified patterns in the syntax literature. Formal assessments of this kind are in a sense similar to the judgment processes which theoretical linguists conduct informally on several examples of the same type. However, introductory textbooks like Adger (2003) intend to present a comprehensive outlook of the syntactic system and typically provide only one or two representative examples for each construction. It is therefore important for us to first ask whether the judgments of those textbook examples are sufficiently robust. As the first step to replicate the judgment data in Chinese syntax, we chose to only focus on example sentences in the book, since they could sometimes be overlooked in studies where statistical results are averaged over multiple items of the same construction. Our decision to include one item per condition was also similar to Sprouse and Almeida (2017) so that the length of the experiment was reasonable for participants to finish without the need of dividing them into several groups. This has also let us to apply the *z*-score transformation to eliminate some forms of rating bias in a task like Likert-scale.[7]

2.2 Participants

148 native speakers of Mandarin Chinese participated in this seven-point Likert scale acceptability rating experiment. They were recruited from social media chat groups in Mainland China. We collected each participant's relevant demographic information, including age, level of education, major of study if currently in college, and profession, as well as Chinese dialects or non-Chinese languages which the participant often uses. Fewer than 5% of our participants received some linguistic training.[8] After completing the approximately 20-minute online experiment, participants were each paid 12 Chinese Yuan for their time.

2.3 Procedure

Experiment 1 was an acceptability rating task administered online using *Qualtrics*, a data collecting website. Sentence items were presented one at a time which participants rated on a scale from 1 (very bad) to 7 (very good) intuitively as a formal measure of linguistic acceptability. The Likert scales have been widely used in the psychological research (Hartley, 2014; Likert, 1932) as well as judgment replication studies.

Two sentences, one acceptable and the other unacceptable, were first introduced as examples. Different from previous experiments, we created two additional catch trials to ensure that the participants paid enough attention throughout the experiment. In each catch trial, participants were directly asked to choose a rate, e.g. the number "2". The information of catch trials was given to participants as a part of the experiment instruction. There was no time limit for rating each sentence and participants were encouraged to rest when needed.

As introduced in Section 2.1, the selected test items belonged to three groups based on whether and how they were compared with other examples, i.e. PAIR, TRIPLE, and OTHER. Table 2 further illustrates how we created five lists of stimuli for Experiment 1

[7] It is also possible to address the rating bias issue by testing baseline items along with target sentences in the same rating experiment. Lin (2018), for example, created three baseline groups in his naturalness-rating experiment by manipulating the degree of word-order and grammaticality violation in sentence items.

[8] Seven participants majored in language-related degree programs, such as linguistics, applied linguistics, Chinese literature, or foreign literature.

Table 2
The five lists of stimuli created for Experiment 1.

List	No. of Test Items			No. of Subjects
	GOOD	QUESTIONABLE	BAD	
1	77	0	81	47
2	81	0	77	51
3	75	7	77	16
4	76	7	76	16
5	78	5	76	18

following the standard practice of cognitive psychology. We divided the 158 PAIR contrasts into two lists, such that for any sentence listed under "List 1" its counterpart within the same contrast will be in "List 2". Since each participant worked on only one list of stimuli, he or she was never exposed to both members of the same PAIR contrast. Three additional lists were then created for sentences in the TRIPLE and OTHER groups. To ensure that the number of stimuli is balanced across all five lists, we also randomly chose and added sentences from the PAIR group to those three lists. Finally, yet importantly, the order of items presented in the experiment was randomized and the sentence type based on their acceptability status was counterbalanced in all lists across participants, in order to avoid possible response bias (Sprouse, 2009).

2.4 Results

After checking the catch trial results, we excluded the data of 18 participants who failed to select the number we directly asked for in any of the two catch trials. The results below are based on rating choices made by the remaining 130 participants. Among them, 98 subjects worked on one of the two PAIR lists while 52 rated sentences in the other three lists.

The mean acceptability ratings for the three sentence types are shown in Figure 1. The overall pattern suggests that unmarked "good" sentences (mean = 5.19) were in general rated higher than "questionable" sentences (mean = 4.22) while "bad" sentences received the lowest ratings (mean = 3.1).

We assessed the statistical significance of overall patterns using two-tailed paired t-tests. The ratings were first standardized within each participant before they were entered into the t-test. The procedure of aggregating Likert scores aimed to mitigate the impact of individual differences among participants with various response styles. We followed the practice of previous studies (Langsford et al., 2018; Linzen & Oseki, 2018; Mahowald et al., 2016; Sprouse & Almeida, 2017; Sprouse et al., 2013) and used z-score transformation for this purpose by subtracting the participant's mean rating and dividing the result by the standard deviation of the participant's ratings. The results in Table 3 show that "good" sentences were rated significantly higher than "questionable" and "bad" sentences. The difference between "questionable" and "bad" sentences was also significant, although it

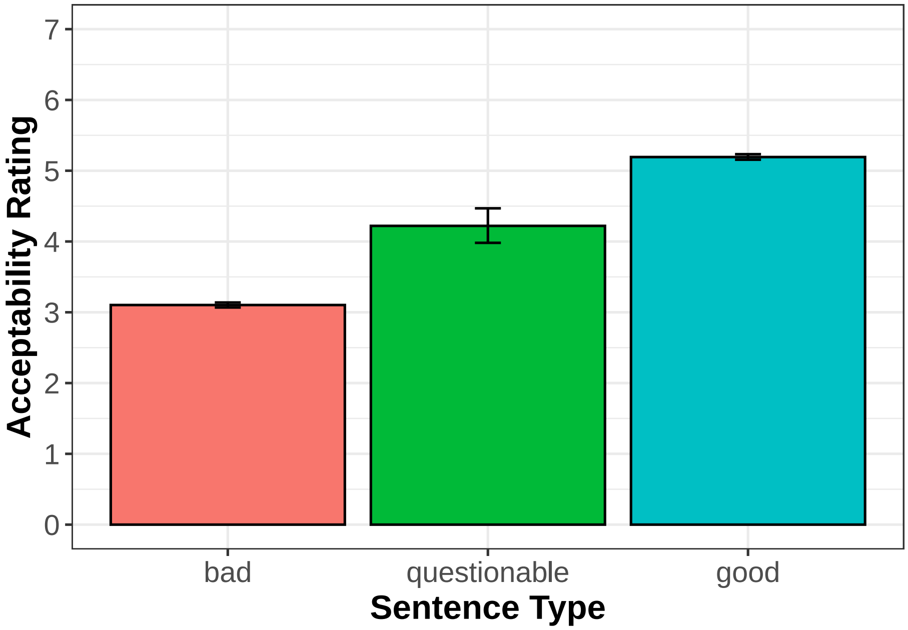

Figure 1. Mean ratings of the three sentence types exhibit the overall acceptability judgment ranking. Error bars represent 95% confidence intervals (CI).

was marginal when the raw data were analyzed (t=2.002).[9]

Table 3
Statistical analyses of overall patterns in Experiment 1.

Contrast	Coefficient	Std. Error	t-value	p
BAD vs QUESTIONABLE	0.43	0.12	3.67	< 0.001
BAD vs GOOD	0.99	0.05	21.87	< 0.0001
QUESTIONABLE vs GOOD	0.55	0.11	4.83	< 0.0001

Next, we turn to the 158 PAIR contrasts, each of which included an unmarked "good" condition and its corresponding "bad" counterpart. As introduced in Table 2, they were further divided and randomized into two lists of stimuli such that a participant would never rate both members from the same contrast.

We examined whether the judgments were stable and therefore reliable over time by checking participants' acceptability ratings in a time series throughout the experiment. Importantly, each circle in Figure 2 represents the mean rating across all "good" or "bad" items presented randomly at a specific time (or at the nth trial) during the test. The results show that participants' acceptability judgments largely remained stable from the first to the last trial and the difference between the "good" and "bad" conditions was not increasingly bigger or smaller. Using a linear mixed-effects model, we evaluated the effect of

[9] In this paper, we choose to also report statistical analyses based on the raw data following recommendations by Juzek (2015) and others.

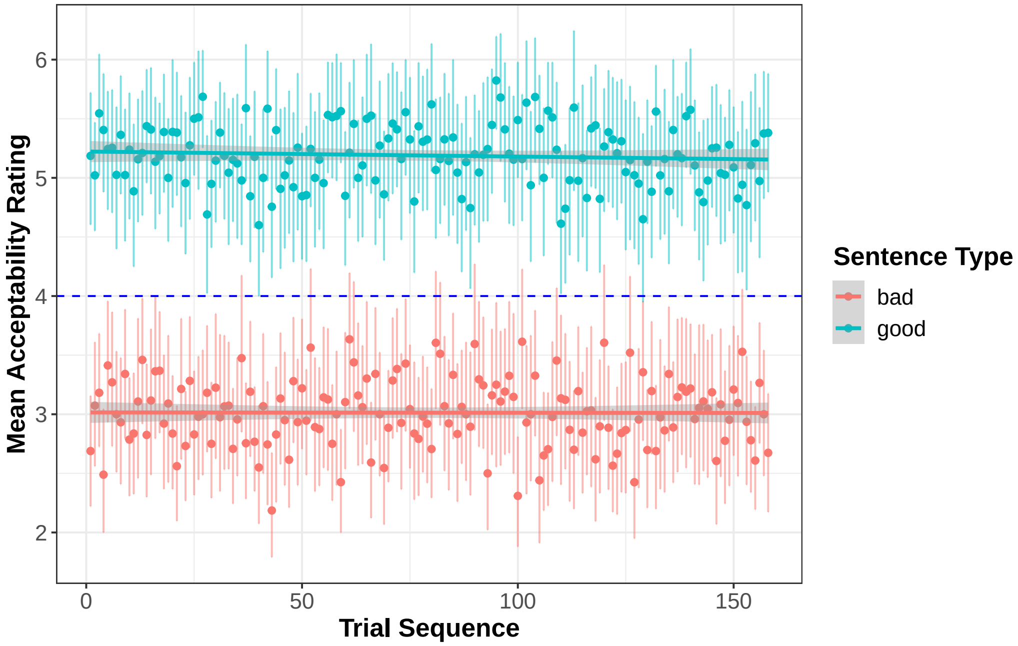

Figure 2. Mean acceptability ratings across all participants indexed in a time series. Each circle represents the mean rating of all "good" or "bad" items at a specific experimental trial (from the 1st to the 158th). Error bars indicate the 95% CI.

trial sequencing, along with three other random variables, i.e. gender, age and education of participants. The results suggest that none of them had a significant impact on how people made rating choices.

Figure 3 shows the distribution of participants' judgment ratings. Each data point represents the acceptability rating which a participant chose for a sentence. In general, the clustering of those data points indicates that "good" sentences tend to get higher ratings than "bad" ones. Indeed, "good" sentences in PAIR contrasts have an mean acceptability rating of 5.19 as compared to 3.01 for "bad" sentences. In this figure, black lines crossing the two columns connect the mean ratings of the two members in a specific PAIR contrast. "Good" sentences were largely rated higher than their counterparts with a few apparent exceptions. The reversed sign of rating difference is indicated by lines trending downward from left to right in Figure 3.

By calculating the numerical rating difference between the two members within the same contrast, we found eight contrasts in which the difference went in the opposite direction than predicted, i.e. a "bad" sentence was rated higher than its counterpart. Those contrasts are illustrated by data points below the dashed blue line in Figure 4.

To evaluate the acceptability rating difference between the "good" and "bad" sentences statistically, we fit a linear mixed model to our data after subjects' ratings were z-transformed. Results suggest that "good" sentences were rated significantly higher than their counterparts within the same contrast ($t = 19.37, p < 0.001$). Applying the same

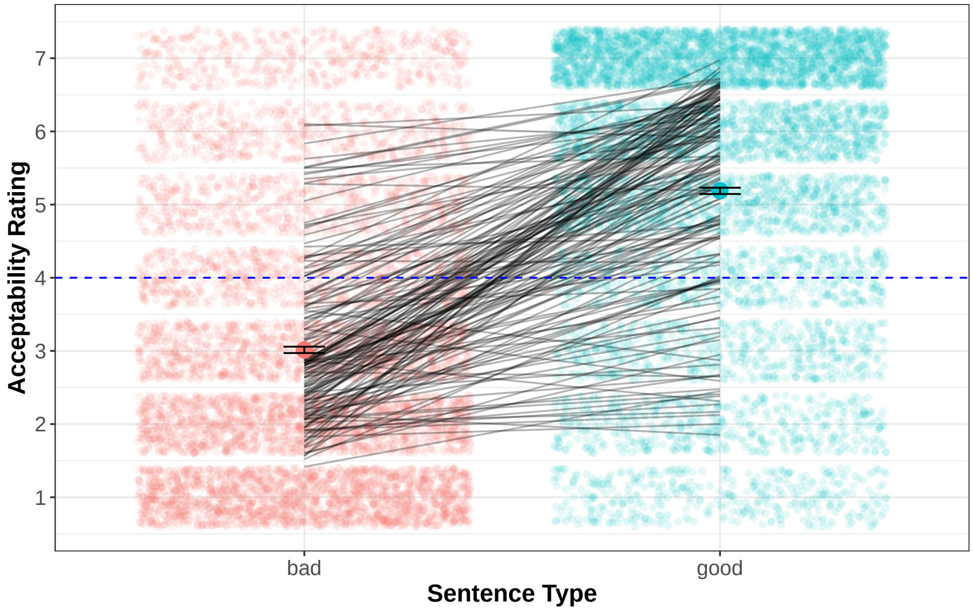

Figure 3. The distribution of acceptability ratings where individual data points are represented by small circles. Blacks lines illustrate the mean rating difference within each PAIR contrast across its two conditions. Mean ratings for the two sentence types are also highlighted with error bars represent the 95% CI.

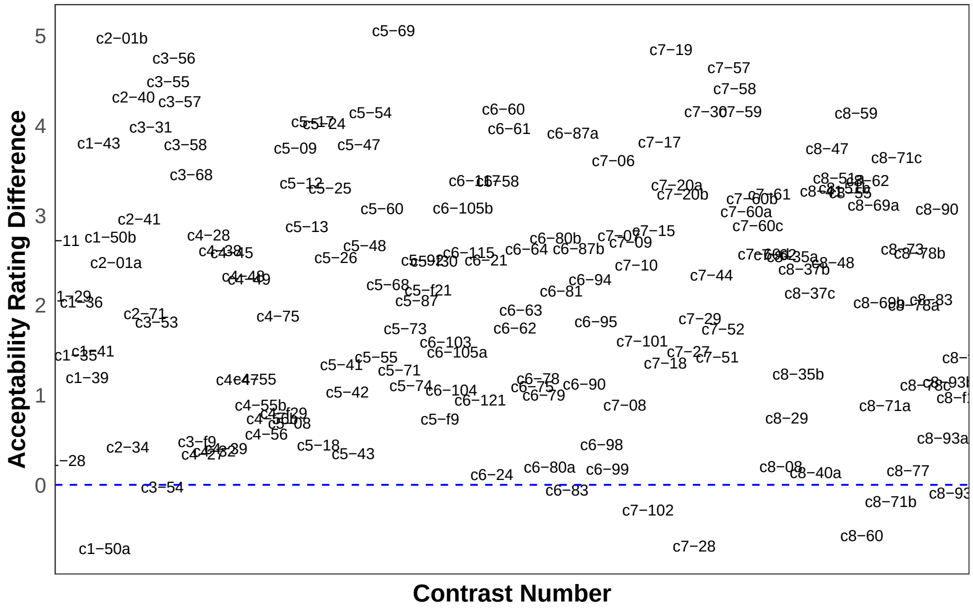

Figure 4. The acceptability rating difference between the two members in a PAIR contrast. Data points are labeled by the identifier number in our experiment.

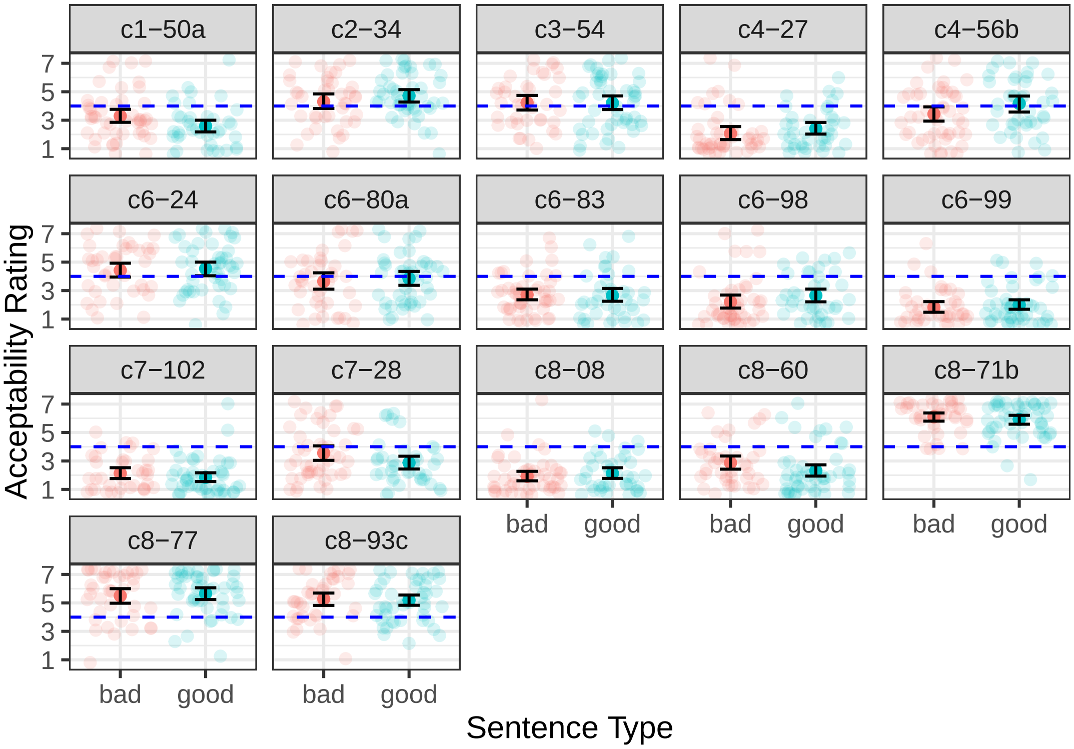

Figure 5. The acceptability rating difference between the two conditions in 17 PAIR contrasts was not significant ($p > 0.05$). Error bars indicate the 95% CI.

method to each individual contrast, 141 of the 158 contrasts replicated at the statistical threshold of $p < 0.05$. The difference did not reach statistical significance in the remaining 17 contrasts, including seven contrasts in which the difference was in the opposite direction than predicted. Table 5 in the Appendix provides a list of those sentence contrasts, along with their example numbers in the book. The distributions of judgments across participants in those contrasts are shown in Figure 5. Fitting the linear mixed model to individual contrasts in the raw data, the rating difference was no longer significant in 9 additional contrasts, as reported in Table 6 and Figure 8 in the Appendix.

In Experiment 1, we replicated the vast majority of judgments in the HLL book using an acceptability rating task. Although there was no clear-cut boundary between the two conditions in 17 PAIR contrasts, the evidence was not sufficient to conclusively determine that those published pairs of examples were indeed problematic. The acceptability judgments can be influenced by a variety of factors in sentence processing, many of which are gradient in nature (Sprouse & Almeida, 2012). Lin (2012), for example, discussed how extra-grammatical processing factors affect the syntactic acceptability of resumptive pronouns in Chinese. In addition, it is possible that "the gradation of acceptability" observed in some of those contrasts are in fact theoretically driven. This notion was first acknowledged in the Subjacency theory literature (Chomsky, 1973) and later realized in the Barriers account (Chomsky, 1986), such that different degrees of acceptability are determined by the number of barriers crossed by movement.[10] Therefore, whether the acceptability ratings alone are informative enough to identify gradient effects predicted by linguistic theories has yet to be examined. Some researchers have discussed this issue with respect to the effect size in confirmatory hypothesis testings. Sprouse and Almeida (2017) argue that, compared to acceptability rating tasks, forced-choice tasks are more robust in detecting such gradient effects.

3 Experiment 2: Forced-choice

3.1 Materials

In Experiment 1, participants rated one sentence at a time on a scale to determine whether it is syntactically acceptable. The judgments were made without directly comparing the sentence with any other examples including its counterpart(s) within the same contrast. Such an "indirect" rating process is somewhat different from how sentence examples are often evaluated and discussed in the syntax literature where researchers focus on the difference between candidates that are of theoretical interest within a particular contrast.[11] To faithfully mirror the common practice of syntactic judgment of critical sentence pairs, in Experiment 2 we adopted the widely used forced-choice task (Myers, 2009a; Rosenbach, 2003; Sprouse et al., 2013). It re-tested all 17 PAIR contrasts identified in the rating task with non-significant difference between the two members. We also created two control groups. The Control 1 group was composed of 27 PAIR contrasts whose rating differences were the most significant in Experiment 1. Those in the Control 2 group were the 9 PAIR contrasts whose rating difference reached significance in the z-transformed data but failed do the same when the raw data were analyzed. The Group 2 sentences can be found in Table 6 of the Appendix.

3.2 Participants

We recruited 86 native speakers of Mandarin Chinese again from social media chat groups, none of whom participated in Experiment 1. Only eight participants were in a language-related major of study. We collected the same set of background data as in Experiment 1 and paid each participant 12 Chinese Yuan after the experiment.

3.3 Procedure

Similar to Experiment 1, we employed *Qualtrics* for this forced-choice task. After two example exercises, participants worked on 53 randomized pairs of sentences including 17 pairs in the Test group, 27 in the Control 1 group, and 9 in the Control 2 group. They simply read and chose the better sentence from each pair. During the experiment, there were again two catch trials in which participants needed to pick the sentence that we directly asked

[10] There is a continuing debate on the gradient acceptability issue in forming syntactic theories (Hofmeister & Sag, 2010; Lau, Clark, & Lappin, 2017; Sprouse, Yankama, Indurkhya, Fong, & Berwick, 2018, among others).

[11] Alternatively, participants can choose from two sentences sampled at random from the set of stimuli. Langsford et al. (2018) compared different acceptability measures, including the RANDOM PAIRS model (Thurstone, 1927) and the TARGET PAIR task used in Experiment 2.

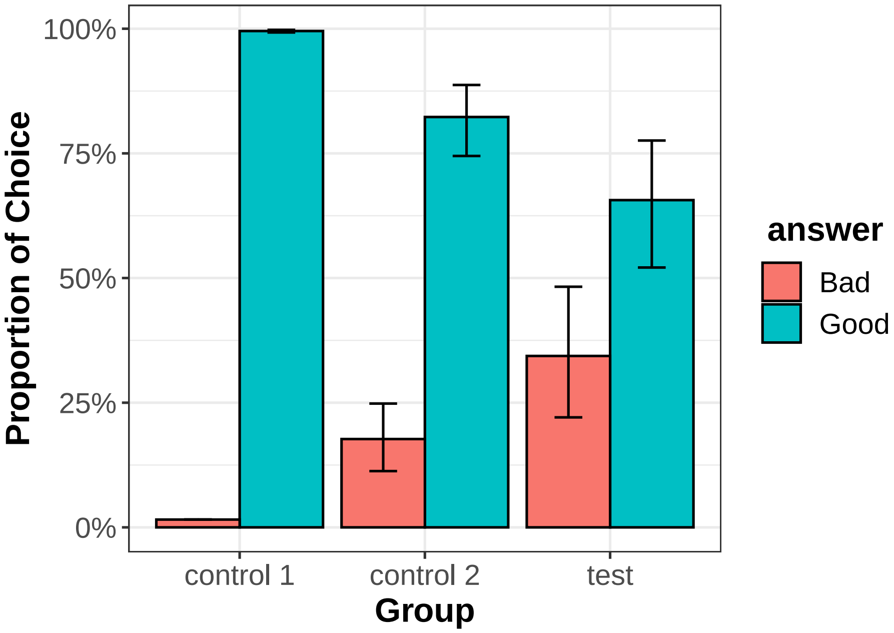

Figure 6. The overall pattern in Experiment 2. Bars represent the proportion of choosing a preferred sentence from a contrast. Error bars indicate the 95% CI.

for. Experiment 2 was not timed and participants were encouraged to take rests during the experiment.

3.4 Results

We first checked how participants performed on the two catch trials. Among all participants, 22 failed to answer at least one of the trials correctly. We excluded their data from further analyses and report the results of the remaining 64 participants below.[12]

As shown in Figure 6, when asked to directly choose the better one from the two sentences in PAIR contrasts, participants picked more "good" sentences across all groups. Specifically, "good" sentences in the Control 1 group were almost always selected. Participants sometimes chose a sentence different from the prediction in the Control 2 group, and more so in the Test group. The overall numerical finding was reconfirmed in Table 4, as we assessed their statistical significance by fitting binominal mixed-effects models to the data. In both control groups and the test group, the proportion of choosing a "good" sentence as the better one was significantly higher.[13]

[12] We suspect that a number of participants may have misunderstood the instruction of catch trials. 15 participants correctly answered at least one catch trial and only seven missed both. Nonetheless, we excluded the data produced by anyone who had failed even one catch trial.

[13] For most contrasts in the Control 1 group, participants did not choose any "bad" sentence. The calculated z-value was therefore smaller in the Control 1 group than in the Control 2 group, as the mixed-effects model considered results of individual contrasts.

Table 4
Statistical analysis results of the overall pattern in Experiment 2.

Group	Estimate	Std. Error	z-value	p
Control 1	6.80	1.16	5.85	< 0.001
Control 2	1.75	0.26	6.68	< 0.001
Test	5.37	0.35	2.33	< 0.05

We looked further into the proportion of choice in each of the PAIR contrasts in the Test group. Experiment 1 found no difference between the two members in those contrasts in terms of acceptability rating. Participants in Experiment 2 had the chance to directly compare the two members in the same contrast in a way that is similar to the introspective judgment process of syntacticians. Indeed, results on those contrasts were more in line with HLL's judgments as compared to the seven-point Likert scale ratings, as shown in Figure 7. After fitting binominal mixed-effects models to each of the contrasts, we found that in the majority of the contrasts the proportion of choosing the "good" sentence as the better one achieved statistical significance ($p < 0.05$). We eventually identified five PAIR contrasts where the participants disagreed with HLL's judgments, as highlighted in Figure 7. The proportional differences were not in the predicted direction in four of those contrasts and there was a numerical advantage of choosing the "good" item in the one remaining (c8-60). Table 5 in the Appendix provides statistical results for each contrast in the Test group. Applying the same analysis to contrasts in the Control 2 group, we found that the proportional difference in all but one contrast (c8-40a) reached significance. Figure 9 and Table 6 in the Appendix report those results.

In all, we used the forced-choice task in Experiment 2 as an alternative method to evaluate the 17 PAIR contrasts which failed to reach significance in the Likert-scale rating experiment. The judgment of native Chinese speakers further converged with the predicted acceptability for published examples in the HLL book. The result were consistent with previous studies comparing various acceptability judgment tasks. Both Sprouse and Almeida (2017) and Langsford et al. (2018) have demonstrated that forced-choice tasks exhibit more sensitivity for detecting the difference between conditions in pairwise contrasts. Together with Experiment 1, the results suggest that the two approaches could be complementary to each other in assessing acceptability judgments of sentence examples in the syntax literature.

4 Discussion

Sharing the same goal with the seminal study by Sprouse and Almeida (2012), this project extend their findings to a non-English language. We conducted two formal experiments on example sentences in the HLL book. Both experiments included a relatively large number of participants that provided sufficient statistical power to detect differences in either acceptability rating or proportion of choices. The results of the seven-point Likert scale rating experiment captured the overall difference between acceptable, questionable, and unacceptable sentences. In particular, 141 of 158 PAIR contrasts replicated, such that

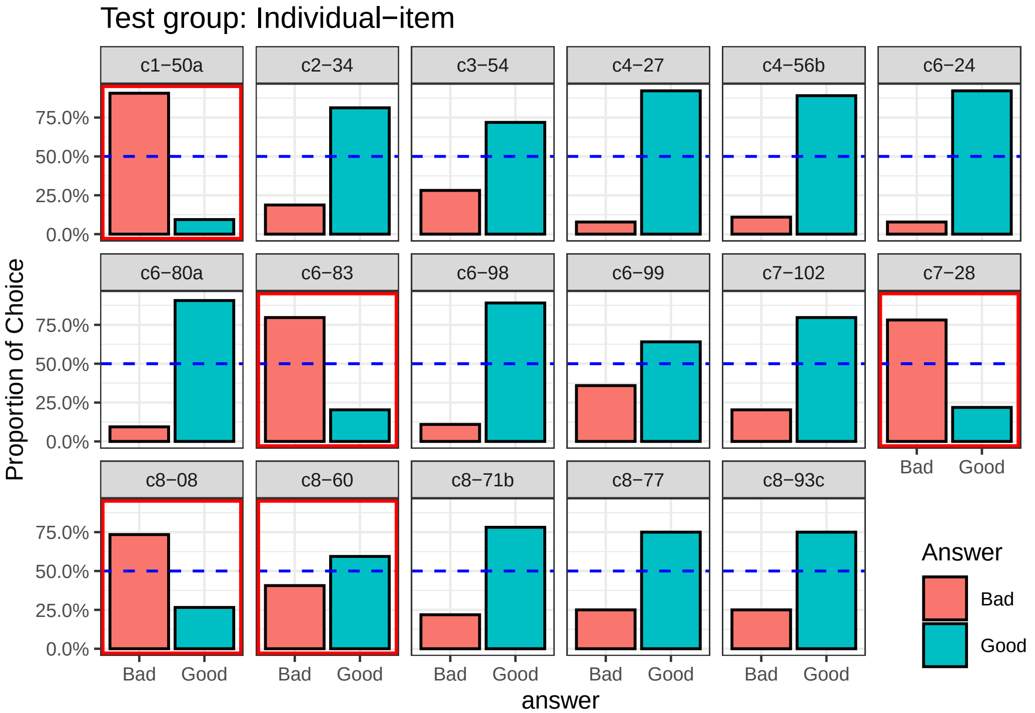

Figure 7. The proportion of choices for contrasts in the Test group. The five contrasts with non-significant difference of choice are highlighted.

an originally reported acceptable sentence was rated significantly higher than its counterpart in the same contrast. We re-tested the 17 non-significant PAIR contrasts in a forced-choice task by asking participants to compare the two conditions directly. The acceptable sentences were more likely to be chosen as the preferred one with 12 more contrasts replicated. Consistent with major findings in Sprouse and Almeida (2012), our results in the two experiments, taken together, provide strong and cross-linguistic evidence, e.g. a 96.8% replication rate of the pairwise contrasts, for the overwhelming reliability of the data in theoretical syntax research.

4.1 A closer look at the contrasts that failed to replicate

As Sprouse and Almeida (2012) argued, "experiments are not truth-discovery machines"' and merely "provide one type of evidence toward a conclusion". Following earlier studies by Sprouse and colleagues, we do not attempt to judge the influence of the five contrasts in question on developing syntactic theories in the book.[14] Instead we simply offer a few possibilities below that may have caused the replication failure on those particular contrasts. The discussion on the quality of theories constructed based on those

[14] We focus on four contrasts where the predicted directionality of results is reversed. For contrast c8-60, more participants indeed chose the acceptable sentence as the preferred one. Its marginal result in Experiment 2 may be a power issue.

data points would not be straightforward because theories are rarely constructed based on a single data point (Linzen & Oseki, 2018; Marantz, 2005; Phillips, 2009; Schütze, 1996; Sprouse & Almeida, 2013, among others). This type of work is best left to theoretical syntacticians who are likely to provide much more informative interpretations in the context of how the examples were used in shaping or rejecting an argument.

Our results, especially in Experiment 1, have shown that the opposite direction (or non-significance) of the acceptability rating difference was primarily because participants rated the "good" sentence in each of the five contrasts much lower than its average rating across the board. In comparison, the ratings of their "bad" counterparts were no different from the unacceptable sentences in other contrasts. We notice that the "good" sentences in three contrasts, re-grouped below in (7a-c), begin with a preposed structure, a typical diagnostic tool in generative syntax research. Sentences with either a preposed prepositional phrase (7a) or a topicalized bare noun phrase (7b and 7c) are legal derivations of the Chinese grammar but their relatively low frequency in the language may have led naïve native speakers to a decision that is different from the HLL's. Therefore, one needs to be cautious about interpreting those results as evidence against the original judgments when factors like the frequency of use are involved.

(7) a. c1-50a; Chapter 1, Example 50(b)
gei wo, ta zuo-le henduo shi.
GEI me he do-LE many thing
'For me, he has done many things.'

b. c6-83, Chapter 6, Example (84)
[[e_i baba] wo kanjian-le de] na-ge nühai$_i$.
father I see-LE DE that–CL girl
'the girl whose father I saw.'

c. c8-08, Chapter 8, Example (10)
xuesheng, wo yiwei chi-le dangao.
student I think eat-LE cake
'The students, I thought (they) ate the cake.'
Not: '(Some) students, I thought (they) ate the cake.'

Additionally, it has long been argued that examples in the syntax literature are sometimes relevant only in relation to an intended meaning, and the preceding context plays an important role in acceptability judgment (Ferreira, 2005; Gibson & Fedorenko, 2013; Newmeyer, 1983; Schütze, 1996). It might be the case in (7c) where the sentence is acceptable only when the topicalized NP *xuesheng* is definite, rather than indefinite or generic, as HLL have argued.[15] Without any preceding context or explanation of the intended meaning, it was challenging for participants in either of the experiments to make an informed or even relevant judgment. Similarly, without specifying a scenario, a global parsing ambiguity arises in (8) when DE is considered as a possessive marker instead of a relativizer, as in *baba de na-ge nühai* "that girl of father's". Participants may find the sentence structurally acceptable based on this additional reading, even though its acceptability is constrained

[15] Bare NPs in Chinese are often ambiguous between definite and indefinite readings.

by semantic plausibility. Subjects were more likely to see it as the better sentence when forced to compare it with its “good” counterpart (7b). A complementary sentence comprehension question could be asked to assist the evaluation of its acceptability under the authors’ intended interpretation.

(8) c6-83, Chapter 6, Example (83)

*[wo kanjian-le [e_i baba] de] na-ge nühai$_i$.
I see-LE father DE that-CL girl

‘the girl that I saw [her] father.’

Lastly, the acceptability of a sentence often hinges on the choice of lexical items. Although Gibson and Fedorenko (2013) made a point that formal experiments with multiple stimuli for each contrast allow linguists to test the robustness of a hypothesis, the controversy over one sentence example should not be considered as an immediate rejection to the syntactic phenomenon it represents. As discussed in Section 2.1, we did not construct and test any additional token other than the HLL sentences. By faithfully assessing only the original textbook examples, we hope to avoid any unintentional bias or mistake that may confound the interpretation of results. This makes our research speak directly to the discussion on the reliability of introspective judgments.

To sum up, we would like to emphasize that intended meaning, context of use, and lexical item choice are a mere few of potentially many factors that can affect the acceptability judgment. It falls outside of the scope of this paper to provide an exhaustive discussion of such factors. We encourage those interested to read Yao, Xie, Lin, and Huang (to appear) for an overview of non-syntactic factors that may impose bearing on the acceptability of Chinese sentences.

4.2 Formal judgment experiments in Chinese

The informal judgments in the syntax literature come under scrutiny at times in the peer-review process or after publication by other linguists, and the studies of Chinese are no exception, including those of HLL’s. For example, Xu (1990, 1996) and Shi (1994) challenged several judgments in Huang (1982) over wh-questions, which HLL’s chapter on Chinese questions was based on. Similarly, disagreements on the judgment of relative clause examples in Aoun and Li (2003) were also reported by Ou (2006). In light of the ongoing debates over informal judgments, there has been a continuing effort to evaluate Chinese sentences quantitatively. In his early work, Myers (2007, 2009a, 2012) took the first step and designed an experimental paradigm to help theoretical syntacticians collect and analyze native speakers’ acceptability judgments on a small scale. He went on to test several controversial topics in Chinese syntax, including islands, topicalization, and number expressions. More recent research roughly along these lines includes contrasting studies on the inverse scope reading in doubly-quantified sentences (Scontras, Polinsky, Tsai, & Mai, 2017; Scontras, Tsai, Mai, & Polinsky, 2014; Zhou & Gao, 2009), an examination of aspect marker selection (Laws & Yuan, 2010), an experiment testing the sensitivity of wh-in-situ questions in complex NP islands (Lu et al., to appear), as well as an investigation on non-canonical classifiers (Gong, Shuai, & Wu, 2019).[16] Those studies have demonstrated the benefits of formal judgment data collection methods in Chinese syntax research.

In a review article, Myers (2009b) argued that "the inconvenience of formal judgment experimentation has been exaggerated, ironically by the reformers themselves, who focus too much on the differences from informal judgments and too little on the similarities". We share his concern and our own work replicated a vast majority of acceptability judgments in the HLL book in a fairly efficient manner. The two experiments in this study each took less than 10 days for completion and were largely unsupervised once the weblinks to the testing site were posted on the Chinese social media. Crowdsourcing platforms, such as Amazon Mechanical Turk, Witmart, and Figure Eight (formerly Crowdflower), can further accelerate the subject recruitment process and gather judgment data comparable to those collected in the laboratory (Sprouse, 2011; Wang, Huang, Yao, & Chan, 2015). Conducting those online experiments has become even more manageable with the help of a new set of free, open-source tools (Erlewine & Kotek, 2016).

It is also worth mentioning that Myers tested whether participants were less sensitive to sentence acceptability over time in parsing Chinese stimuli (Myers, 2007, 2009a). This kind of satiation effect has been discussed extensively in the sentence processing literature of syntactic islands where unacceptable sentences become more acceptable in some constructions (Chaves & Dery, 2014; Do & Kaiser, 2017; Francom, 2009; Goodall, 2011; Hiramatsu, 2000; Snyder, 2000; Sprouse, 2009). In Figure 2, we have seen that participants' ratings remained stable over time in the acceptability rating experiment. The statistical analysis found no effect of trial sequencing as a random variable. It gave us confidence in participants' linguistic intuitions even after they were repeatedly tested. On the other hand, because our study assessed fully randomized judgments on a large variety of syntactic phenomena, rather than multiple items on a single topic, the results should not be seen as counter-evidence against the satiation effect, which mostly focused on certain types of island constructions.

4.3 Dialectal and generational variations

In their response to Featherston (2007), den Dikken et al. (2007) raised important questions about the impact of dialectal variation in the population in formal replication studies. Experimentalists have to sample from a correctly defined population of participants, regardless of the recruitment method. Indeed, the acceptability of Chinese sentence examples may vary greatly due to speakers' dialectal differences, as HLL have also acknowledged in their book. The participants we recruited were mostly students in Chinese national universities with a diverse geographic background. Although we collected their information on speaking a dialect or a language other than Mandarin Chinese, those self-reported data were not included as a random variable in our analyses because they require further verification, including the speaker's proficiency level of the dialect and the frequency of using it. Nonetheless, we have found little evidence from our results of patterns that could be ascribed to different dialects. But the framework of this study makes it possible for future works to experimentally examine syntactic variations in Chinese spoken

[16] See Fukuda (to appear) for a comprehensive review of studies using acceptability and truth value judgment methods in East Asian languages, including Chinese.

across different regions, in addition to large-scale corpus studies (Khoo & Lin, 2018).

Similarly, there may be generational differences in syntactic judgments between the authors and our participants as a group, especially as the Chinese language continues to evolve. A large number of examples in the HLL book were taken from the authors' publications in the 1980s and 1990s. In contrast, most participants in our experiments were graduate students with an average age of 23.9. Although the generational question cannot be addressed by our data, it would be interesting to compare the judgments of native speakers in different age groups in a future experiment.

5 Conclusion

Acceptability judgments of sentence examples are crucial in forming syntactic theories. Recent methodological debates over introspective judgments made by individual linguists have led to a series of experiments replicating data points in published textbooks and peer-reviewed journal articles. Sprouse and colleagues, in particular, have shown that native speakers agree with most linguists' English judgments. Although the replication rates reported are lower in previous studies of Hebrew, Korean, and Japanese, whether the issue of judgment reliability is amplified in languages other than English requires further investigation.

In this paper, we empirically assessed this concern by formally testing 557 data points in the popular textbook, Huang et al. (2009), including 158 pairs of sentences with contrastive acceptability. To our knowledge, this was the first attempt to replicate judgments in a non-English generative syntax textbook. The statistical analyses of two large-scale experiments (acceptability rating and forced-choice) suggest that the vast majority of Chinese judgments in the book are robust. Only five acceptability contrasts failed to replicate in both tasks. As the first attempt to examine Chinese judgments on a large-scale, we implemented a practical experimental framework in the hope of bridging the gap between the informal data-collection in Chinese linguistic research and the protocols of experimental cognitive science. By covering sentence examples within a wide range of phenomena in Chinese syntax, this work provides informative judgments by naïve Chinese speakers and establishes a quantitative baseline for future research.

Although this paper can certainly be read alongside Sprouse and Almeida (2012) as both studies have sought to answer a similar research question, our scopes, focuses and experimental manipulations are not entirely the same. The most obvious one, as we have mentioned previously, is that we only assessed the published examples in the HLL book without adding more self-constructed items for each phenomenon. We believe that testing multiple items for the same construction is valuable in determining the generalizability of a particular syntactic phenomenon. Similarly, there is also a need for comparing multiple criteria and statistical tools to show stronger evidence of replicability, which has been discussed in great details in Schütze and Sprouse (2014). Finally, one natural future direction of this project is to explore appropriate experimental methods to assess the excluded data points. Juzek and Häussler (to appear), for example, have recently emphasized the importance of replicating sentence examples that are independent of counterparts. As for sentences with co-referential dependencies, it might also be possible to evaluate their acceptability by creating a preceding context or a comprehension question explicitly asking about the reader's choice of antecedent (Chen, Jäger, & Vasishth, 2012).

Acknowledgements

This work was supported by a faculty research fund to Zhong Chen from the College of Liberal Arts at Rochester Institute of Technology. We are grateful to Jeff Runner for discussions at various stages of this project. We thank Tian Tian for her assistance in preparing experimental materials, Qingrong Chen, Qiongpeng Luo and Zhuang Wu for helping with recruiting participants, as well as Jacquelyn Haller for editorial suggestions. We would also like to thank the anonymous reviewers and the editors of JEAL for their insightful comments and suggestions.

References

Adger, D. (2003). *Core syntax: A minimalist approach*. Oxford University Press.

Aoun, J., & Li, Y.-h. A. (2003). *Essays on the representational and derivational nature of grammar: The diversity of wh-constructions*. MIT Press.

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence*. Dordrecht: Springer.

Chaves, R. P., & Dery, J. E. (2014). Which subject islands will the acceptability of improve with repeated exposure? In R. E. Santana-LaBarge (Ed.), *Proceedings of the 31st West Coast Conference on Formal Linguistics.* Somerville, MA: Cascadilla Proceedings Project.

Chen, Z., Jäger, L., & Vasishth, S. (2012). How structure-sensitive is the parser? Evidence from Mandarin Chinese. In B. Stolterfoht & S. Featherston (Eds.), *Empirical approaches to linguistic theory* (pp. 43–62). Studies in Generative Grammar, Berlin: Mouton de Gruyter.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1973). Conditions on transformations. In S. Anderson & P. Kiparsky (Eds.), *A Festschrift for Morris Halle* (pp. 232–286). New York: Holt, Reinhart and Winston.

Chomsky, N. (1986). *Barriers*. MIT press.

Cowart, W. (1997). *Experimental syntax: Applying objective methods to sentence judgements*. Thousand Oaks, CA: SAGE Publications.

den Dikken, M., Bernstein, J. B., Tortora, C., & Zanuttini, R. (2007). Data and grammar: Means and individuals. *Theoretical Linguistics*, *33*(3), 335–352.

Do, M. L., & Kaiser, E. (2017). The relationship between syntactic satiation and syntactic priming: A first look. *Frontiers in Psychology: Language Sciences*, *8*(1851).

Edelman, S., & Christiansen, M. H. (2003). How seriously should we take minimalist syntax? A comment on Lasnik. *Trends in Cognitive Sciences*, *7*(2), 60–61.

Erlewine, M. Y., & Kotek, H. (2016). A streamlined approach to online linguistic surveys. *Natural Language & Linguistic Theory*, *34*(2), 481–495.

Featherston, S. (2005). Magnitude estimation and what it can do for your syntax: Some wh-constraints in German. *Lingua*, *115*(11), 1525–1550.

Featherston, S. (2007). Data in generative grammar: The stick and the carrot. *Theoretical linguistics*, *33*(3), 269–318.

Ferreira, F. (2005). Psycholinguistics, formal grammars, and cognitive science. *The Linguistic Review*, *22*(2-4), 365–380.

Francom, J. C. (2009). *Experimental syntax: Exploring the effect of repeated exposure to anomalous syntactic structure–evidence from rating and reading tasks* (Unpublished doctoral dissertation).

Fukuda, S. (to appear). Acceptability and truth value judgment studies in East Asian languages. In G. Goodall (Ed.), *The Cambridge Handbook of Experimental Syntax.* Cambridge University Press.

Gibson, E., & Fedorenko, E. (2010). Weak quantitative standards in linguistics research. *Trends in Cognitive Sciences*, *14*(6), 233–234.

Gibson, E., & Fedorenko, E. (2013). The need for quantitative methods in syntax and semantics research. *Language and Cognitive Processes*, *28*(1-2), 88–124.

Gibson, E., Piantadosi, S. T., & Fedorenko, E. (2013). Quantitative methods in syntax/semantics research: A response to Sprouse and Almeida (2013). *Language and Cognitive Processes*, *28*(3), 229–240.

Gong, T., Shuai, L., & Wu, Y. (2019). The acceptability judgment of Chinese pseudo-modifiers with and without a sentential context. *PLOS ONE*, *14*(7).

Goodall, G. (2011). Syntactic satiation and the inversion effect in English and Spanish wh-questions. *Syntax*, *14*(1), 29–47.

Hartley, J. (2014). Some thoughts on Likert-type scales. *International Journal of Clinical and Health Psychology*, *14*(1), 83–86.

Hiramatsu, K. (2000). *Accessing linguistic competence: Evidence from children's and adults' acceptability* *judgments* (Unpublished doctoral dissertation). University of Connecticut.

Hofmeister, P., & Sag, I. A. (2010). Cognitive constraints and island effects. *Language*, *86*(2), 366–415.

Huang, C.-T. J. (1982). *Logical relations in Chinese and the theory of grammar* (Unpublished doctoral dissertation). MIT.

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2009). *The Syntax of Chinese*. Cambridge University Press.

Huang, C.-T. J., Li, Y.-H. A., & Li, Y. (2013). *Hanyu Jufa Xue* 汉语句法学 *[The Syntax of Chinese]* (Simplified Chinese ed.; Y. Gu, Ed. & H. Zhang, Trans.). Beijing, China: World Publishing Corporation.

Juzek, T. (2015). *Acceptability judgement tasks and grammatical theory* (Unpublished doctoral dissertation). University of Oxford.

Juzek, T., & Häussler, J. (to appear). Data convergence in syntactic theory and the role of sentence pairs. *Zeitschrift für Sprachwissenschaft*.

Keller, F. (2000). *Gradience in grammar: Experimental and computational aspects of degrees of grammaticality* (Unpublished doctoral dissertation). University of Edinburgh.

Khoo, Y. K., & Lin, J. (2018). Grammatical variations between Singapore, Mainland China, and Taiwan Mandarin: A pilot study of aspect marking. In *Proceedings of the 32nd Pacific Asia conference on language, information and computation.*

Labov, W. (1978). Sociolinguistics. In W. O. Dingwall (Ed.), *A survey of linguistic science* (pp. 339–72). Stamford, CT: Greylock.

Langendoen, D. T., Kalish-Landon, N., & Dore, J. (1973). Dative questions: A study in the relation of acceptability to grammaticality of an English sentence type. *Cognition*, *2*(4), 451–478.

Langsford, S., Perfors, A., Hendrickson, A. T., Kennedy, L. A., & Navarro, D. J. (2018). Quantifying sentence acceptability measures: Reliability, bias, and variability. *Glossa: a journal of general linguistics*, *3(1)*(37), 1–34.

Lau, J. H., Clark, A., & Lappin, S. (2017). Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge. *Cognitive Science*, *41*(5), 1202–1241.

Laws, J., & Yuan, B. (2010). Is the core-peripheral distinction for unaccusative verbs cross-linguistically consistent?: Empirical evidence from Mandarin. *Chinese Language and Discourse*, *1*(2), 220–263.

Levelt, W. J. M., van Gent, J. A. W. M., Haans, A. F. J., & Meijers, A. J. A. (1977). Grammaticality, paraphrase, and imagery. In S. Greenbaum (Ed.), *Acceptability in language* (pp. 87–101). Mouton The Hague.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*(140), 44-60.

Lin, C.-J. C. (2012). Distinguishing grammatical and processing explanations of syntactic acceptability. In J. Myers (Ed.), *In search of grammar: Experimental and corpus-based studies.* Language and Linguistics Monograph Series (Vol. 48), Academia Sinica, Taipei.

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 journal of general linguistics*, *3(1)*(100), 1–25.

Lu, J., Thompson, C. K., & Yoshida, M. (to appear). Chinese wh-in-situ and islands: A formal judgment study. *Linguistic Inquiry*.

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.

Marantz, A. (2005). Generative linguistics within the cognitive neuroscience of language. *The Linguistic Review*, *22*(2-4), 429–445.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., . . . Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In *Proceedings of the NAACL-HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical*

Turk (pp. 122–130).

Myers, J. (2007). MiniJudge: Software for small-scale experimental syntax. In *International Journal of Computational Linguistics & Chinese Language Processing* (Vol. 12, pp. 175–194).

Myers, J. (2009a). The design and analysis of small-scale syntactic judgment experiments. *Lingua*, *119*(3), 425–444.

Myers, J. (2009b). Syntactic judgment experiments. *Language and Linguistics Compass*, *3*(1), 406–423.

Myers, J. (2012). Testing adjunct and conjunct island constraints in Chinese. *Language and Linguistics*, *13*(3), 437.

Newmeyer, F. J. (1983). *Grammatical theory: Its limits and its possibilities*. University of Chicago Press.

Newmeyer, F. J. (2013). Goals and methods of generative syntax. In M. den Dikken (Ed.), *The Cambridge handbook of generative syntax* (pp. 61–92). Cambridge University Press Cambridge, UK.

Ou, T.-S. (2006). *Suo relative clauses in Mandarin Chinese* (Unpublished master's thesis). National Chung Cheng University, Taiwan.

Phillips, C. (2009). Should we impeach armchair linguists? In S. Iwasaki (Ed.), *Japanese/korean linguistics* (Vol. 17, pp. 49–64). CSLI Publications.

Phillips, C., & Lasnik, H. (2003). Linguistics and empirical evidence: Reply to Edelman and Christiansen. *Trends in Cognitive Sciences*, *7*(2), 61–62.

Rosenbach, A. (2003). Aspects of iconicity and economy in the choice between the s-genitive and the of-genitive in English. In G. Rohdenburg & B. Mondorf (Eds.), *Determinants of grammatical variation in english* (Vol. 43, p. 379-412). De Gruyter Mouton.

Schütze, C. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistic methodology*. University of Chicago Press.

Schütze, C. (2020). Acceptability ratings cannot be taken at face value. In S. Schindler, A. Drozdzowicz, & K. Brøcker (Eds.), *Linguistic intuitions: Evidence and method* (chap. 11). Oxford University Press.

Schütze, C., & Sprouse, J. (2014). Judgment data. In R. J. Podesva & D. Sharma (Eds.), *Research methods in linguistics* (p. 27-50). Cambridge University Press.

Scontras, G., Polinsky, M., Tsai, C.-Y. E., & Mai, K. (2017). Cross-linguistic scope ambiguity: When two systems meet. *Glossa: a journal of general linguistics*, *2*(1), 1–28.

Scontras, G., Tsai, C.-Y. E., Mai, K., & Polinsky, M. (2014). Chinese scope: An experimental investigation. In *Proceedings of Sinn und Bedeutung* (Vol. 18, pp. 396–414).

Shi, D. (1994). The nature of Chinese wh-questions. *Natural Language & Linguistic Theory*, *12*(2), 301–333.

Snyder, W. (2000). An experimental investigation of syntactic satiation effects. *Linguistic Inquiry*, *31*(3), 575–582.

Song, S., Choe, J.-W., & Oh, E. (2014). FAQ: Do non-linguists share the same intuition as linguists? *Language Research*, *50*(2), 357–386.

Sprouse, J. (2009). Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry*, *40*(2), 329–341.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior research methods*, *43*(1), 155–167.

Sprouse, J., & Almeida, D. (2012). Assessing the reliability of textbook data in syntax: Adger's *Core Syntax*. *Journal of Linguistics*, *48*(3), 609–652.

Sprouse, J., & Almeida, D. (2013). The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*, *28*(3), 222–228.

Sprouse, J., & Almeida, D. (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa: a journal of general linguistics*, *2(1)*(14), 1–32.

Sprouse, J., Schütze, C., & Almeida, D. (2013). Assessing the reliability of journal data in syntax: *Linguistic Inquiry* 2001–2010. *Lingua*, *134*, 219–248.

Sprouse, J., Wagers, M., & Phillips, C. (2012). Working-memory capacity and island effects: A reminder of the issues and the facts. *Language*, *88*(2), 401–407.

Sprouse, J., Yankama, B., Indurkhya, S., Fong, S., & Berwick, R. C. (2018). Colorless green ideas do sleep furiously: Gradient acceptability and the nature of the grammar. *The Linguistic Review*, *35*(3), 575-599.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review*, *34*(4), 273.

Wang, S., Huang, C.-R., Yao, Y., & Chan, A. (2015). Mechanical turk-based experiment vs laboratory-based experiment: A case study on the comparison of semantic transparency rating data. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation* (pp. 53–62).

Xu, L. (1990). Remarks on lf movement in chinese questions. *Linguistics*, *28*(2), 355–383.

Xu, L. (1996). Construction and destruction of theories by data: A case study. In *Chicago linguistics society* (Vol. 32, pp. 107–118).

Yao, Y., Xie, Z., Lin, C.-J. C., & Huang, C.-R. (to appear). Acceptability or Grammaticality: Judging Chinese Sentences for Linguistic Studies. In *Cambridge handbook of Chinese linguistics.* Cambridge University Press.

Zhou, P., & Gao, L. (2009). Scope processing in Chinese. *Journal of Psycholinguistic Research*, *38*, 11-24.

Appendix

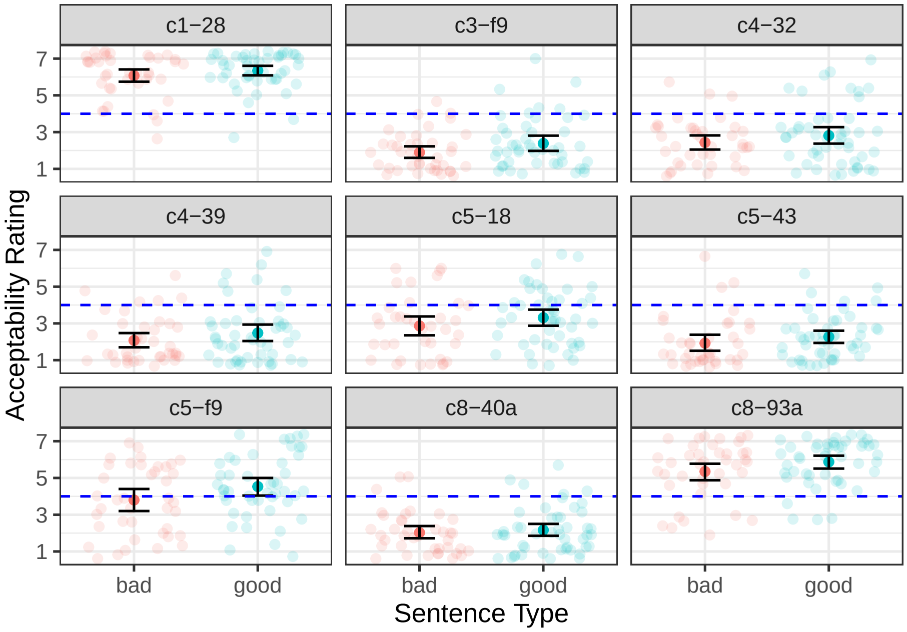

Figure 8. Analyzing the raw data of Experiment 1 suggests that the rating difference was not significant between the two conditions in 9 PAIR contrasts, in addition to those in Figure 5. Error bars represent the 95% CI.

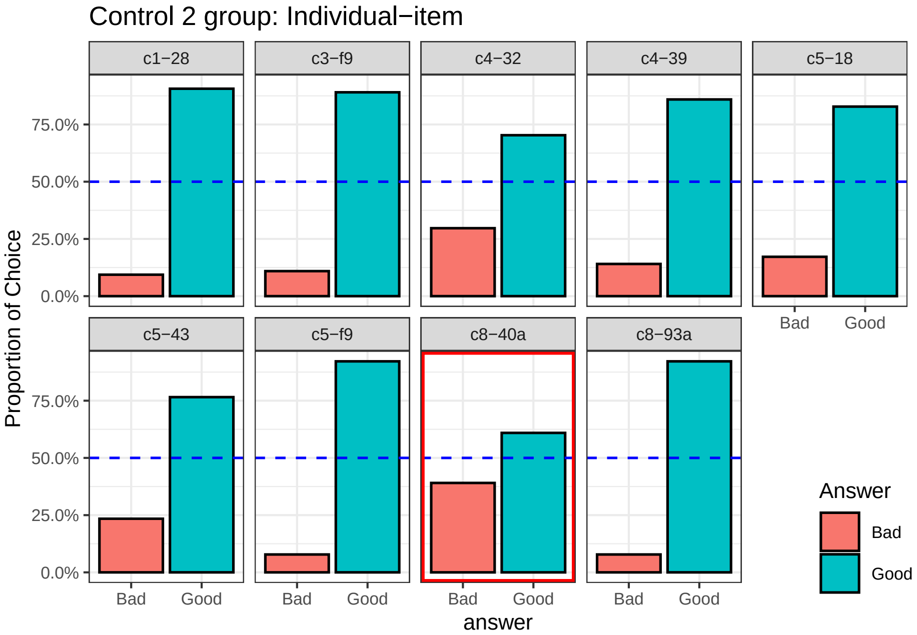

Figure 9. The proportional difference of choosing a "good" or "bad" sentence was significant in all but one contrast in the Control 2 group of Experiment 2.

Table 5

The acceptability rating differences between the two members were not significant in 17 PAIR *contrasts* ($t > 2$) *when the* z*-transformed data were analyzed in Experiment 1. Re-testing them in Experiment 2 using forced-choice suggests that 5 contrasts, highlighted in red, again failed to reach significance statistically* ($z > 2$).

Chapter	Identifier	No.	Sentence	Experiment 1 Mean Rating	 t	Experiment 2 Prop. of Choice	 z
1	c1-50a	50b	给我，他做了很多事。	2.59	-3.67	9%	-5.29
		50d	*他做了很多事给我。	3.30		91%	
2	c2-34	34a	这艘摩托艇已经被小心翼翼地开了许多年了。	4.72	1.94	81%	4.58
		34b	??这艘摩托艇已经小心翼翼地开了许多年了。	4.30		19%	
3	c3-54	54a	我没悄悄地回家。	4.21	0.77	72%	3.37
		54b	*我悄悄地没有回家。	4.23		28%	
4	c4-27	27	张三被我通知李四把赞美他的书都买走了。	2.41	-0.18	92%	5.30
		27	*张三被我通知李四把赞美的书都买走了。	2.06		8%	
	c4-56b	55d	我们学校被警察逮捕了两个学生。	4.17	1.16	89%	5.24
		56d	*我们学校被警察捕了两个学生。	3.44		11%	
6	c6-24	25	水果，我最喜欢香蕉。	4.54	1.22	92%	5.30
		26	*张三，我最喜欢爸爸。	4.42		8%	
	c6-80a	80a	李四看到张三的地方。	3.85	1.20	91%	5.29
		80a	*李四所看到张三的地方。	3.65		9%	
	c6-83	84	爸爸我看见了的那个女孩。	2.65	-1.07	20%	-4.40
		83	*我看见了爸爸的那个女孩。	2.71		80%	
	c6-98	99a	我想看你说每个人会带回来的自己的朋友。	2.66	0.47	89%	5.24
		98a	*我想看你说每个人会带他回来的自己的朋友。	2.21		11%	

	c6-99	99b	我想看你说每个人会带回来的我介绍过给他的朋友。	2.00	1.98	64%	2.22
		98b	*我想看你说每个人会带他回来的我介绍过给他的朋友。	1.82		36%	
7	c7-28	28a	到底有一个人买了什么？	2.87	-2.82	22%	-4.21
		28b	*有一个人到底买了什么？	3.55		78%	
	c7-102	102b	张三想知道什么李四买了。	1.85	0.29	80%	4.40
		102a	*什么张三想知道李四买了。	2.13		20%	
8	c8-08	10	学生，我以为吃了蛋糕。	2.12	-0.61	27%	-3.59
		8a	*三个学生，我以为吃了蛋糕。	1.92		73%	
	c8-60	59b	我对小强们三个特别好。	2.31	-1.13	59%	1.49
		59d	*我对三个小强们特别好。	2.88		41%	
	c8-71b	71b	两个人可以吃十碗饭。	5.92	-0.02	78%	4.21
		71b	*两个人吃十碗饭。	6.1		22%	
	c8-77	77b	如果一只大象鼻子很长，那一定很可爱。	5.66	0.94	75%	3.81
		77a	*一只大象鼻子很长。	5.50		25%	
	c8-93c	93c	要是有一个人很有钱，我们就去找他资助。	5.19	0.19	72%	5.30
		93c	*要是一个人很有钱，我们就去找他资助。	5.28		25%	

Table 6
In Experiment 2, the proportional differences of choice in all but one contrast in the Control 2 group reached significance.

Chapter	Identifier	No.	Sentence	Experiment 1	Experiment 2	
				Avg. Rating	Prop. of Choice	z
1	c1-28	28a	他对这个结局很不满。	6.35	91%	5.29
		28b	??他很不满这个结局。	6.08	9%	
3	c3-f9	f9ii	他小曲没有唱过。	2.38	89%	5.24
		f9iii	*他没有小曲唱过。	1.9	11%	
4	c4-32	33	李四打了他一下的那个人来了。	2.81	70%	3.15
		32	??李四打了他的那个人来了。	2.44	30%	
	c4-39	39	这件事跟他没有关系的那个人走了。	2.48	88%	5.03
		39	*这件事跟没有关系的那个人走了。	2.08	12%	
5	c5-f9	f9a	我抽烟，第一口就把我呛得连连咳嗽。	4.53	92%	5.30
		f9b	*我抽烟，就被第一口呛得连连咳嗽。	3.8	8%	
	c5-18	18a	林伊又被王五击出了一支全垒打。	3.31	83%	4.75
		18b	??王五又把 林伊击出了一支全垒打。	2.86	17%	
	c5-43	43a	张三使我打伤他。	2.27	77%	4.01
		43b	*张三把我打伤他。	1.92	23%	
8	c8-40a	40a	我看到过一个小明。	2.17	61%	1.74
		40b	*我看到过一个小明那个糊涂蛋。	2.03	39%	
	c8-93a	93a	如果有一个人在等他，他就得马上回去。	5.87	92%	5.30
		93a	*如果一个人在等他，他就得马上回去。	5.35	8%	