Experimental Design with a Focus on Power Calculations

Masashi Harada

July, 2020

1 Introduction

This paper describes the design of an acceptability judgment experiment with a particular focus on the power calculation. The experiment examines whether there is a significant difference in acceptability between a particular type of Japanese sentences uttered in two different types of contexts. While the author found in his informal data collection a tendency that the sentences uttered in different types of contexts have different acceptability, some individual variabilities were observed. Thus, it is necessary to conduct a formal experiment and collect data from a larger sample to make any inference or conclusion.

In designing the experiment, one of the main questions concerns how large the sample size of this experiment should be in order to detect the difference, if any, with enough power, which this paper assume to be 80% following the statistical convention. Due to the scarce of literature on the Japanese phenomenon in question or literature on the statistics about the acceptability judgment of Japanese sentences, it is difficult to estimate the accurate sample size required for the experiment. However, this paper attempts to estimate it by doing simulation with a created data frame and linear mixed-effects model. The simulation will be based on the acceptability judgment data from Linzen and Oseki's (2018) experiment about a Japanese phenomenon, which seems to be the most relevant to the phenomenon to be examined. The simulation will prevent estimating too small sample size for the experiment and increses the likelihood of avoiding Type II errors.

The organization of this paper is as follows. Section 2 first briefly explains the phenomenon the experiment is interested in. Section 3 then overviews the design of the experiment. In Section 4, I present the power analysis, based on which the number of required participants is estimated in Section 3. Finally, Section 5 concludes.

2 Phenomenon to be examined

The experiment investigates a phenomenon that the predicate in Japanese copular constructions as in (1) can sometimes hold accusative case but sometimes cannot.

(1) kyoo-wa onigiri-**o** mit-tu dayo
today-Topic rice.ball-Accusative 3-Classifier copula
'Today is three rice balls'

In (1), the predicate *onigiri-o mit-tu* 'three rice balls' can optionally hold accusative case *-o*. The availability of the accusative case depends on the utterance context where the sentence occurs. For example, sentence (1) is natural in the context in (2a) but not in the context in (2b).

(2) a. Ken is the father of Ai, and always cooks lunch for her. It is 6am now. Ai has just come to kitchen, and Ken says (1) to Ai.

b. Ken and Ai have long been examining when different kinds of food they put in a showcase goes bad. Ken always checks which food and how many of them have gone bad. It is 4pm. Ai has just come to the showcase. Looking at the food, Ken says (1) to Ai.

The contextual variability in the acceptability of the accusative case raises various descriptive and theoretical questions. Among those, the one that is relevant here concerns what kind of contexts allow the accusative case. As an answer to this question, Harada (2018) puts forward a descriptive generalization in the same line with (3).

(3) The predicate accusative case in Japanese copular sentences is available only when the context supports accommodation of a question which:

a. if expressed linguistically, contains an accusative case-marked wh-item, and

b. the copular sentence answers.

I call wh-questions satisfying the conditions in (3) wh_{Acc}-question.

I demonstrate that whereas the context in (2a) accommodates a wh_{Acc}-question, the one in (2b) does not, and thus only the context in (2a) allows the predicate accusative case. First, consider a wh_{Acc}-question accommodated in (2a) below.

(4) Ken-wa **nani-o** tukutta-no?
Ken-Topic.marker what-Accusative made-Question.marker
'What did Ken make?'

The question with an accusative case-marked wh-item in (4) is contextually salient because Ken always cooks lunch for Ai every morning and (1) is uttered in the morning. Also, the copular sentence in (1) answers the question in (4). Thus, the question in (4) is an wh_{Acc}-question accommodated in (2a).

In contrast to (2a), it is difficult to envision a wh_{Acc}-question in context (2b); the most natural wh-question to accommodate in (2b) that is answered by (1) would be (5). But the question does not contain an accusative case-marked wh-item. Thus, (5) is not a wh_{Acc}-question.

(5) kyoo-wa **nani-ga/*o** kusatta?
today-Topic.marker what-Nominative/Accusative went.bad
'What has gone bad today?'

To sum up, the availability of the predicate accusative case depends on the utterance context, and seems to be governed by the conditions in (3). However, as mentioned in Section 1, there are some differences in acceptability among individuals. Therefore, the proposed experiment examines whether (3) is a correct generalization of when the predicate accusative case is available.

3 Overview of the experiment

3.1 Participants

I will recruit at the very least 48 native speakers of Japanese for the experiment.[1] I will recruit them in a platform called *Crowdworks*. Following Linzen and Oseki (2018), I will recruit only Japanese speakers who satisfy the following two conditions: (i) they lived in Japan from birth until (at least) age 13, and (ii) their parents spoke Japanese to them at home. While speakers of different dialects of Japanese potentially show different patterns in acceptability judgment, it is not clear whether this is indeed the case. So I will not prevent speakers of any dialects of Japanese from participating in the experiment. But I will ask participants to claim their dialects in the questionnaire so I can examine the difference in the judgment among different dialect speakers. Likewise, I will allow both linguists and non-linguists to participate in the experiment. While there are some empirical evidence that those two populations may provide different judgments (e.g., Spencer 1973, Gordon and Hendrick 1997, Dabrowska 2010), it still remains unclear whether linguists should not collect elicit data from lingusits (e.g., Schütze and Sprouse 2014); whereas linguists's judgments may be affected by their theoretical viewpoints (e.g., Edelman and Christiansen 2003, Wasow and Arnold 2005, Gibson and Fedorenko 2013), linguists may be more sensitive to subtle differences in acceptability that non-linguists may fail to detect (e.g., Newmeyer 1983, Grewendorf 2007). But I will ask participants in the questionnaire whether they have learned linguistics.

3.2 Materials

The experiment will invovle 16 main items, each of which consists of the same sequence of Japanese copular sentences uttered in two different contexts. Thus, there are 32 main sentences in the experiment. One of the example items consists of sentence (1) in context (2a) and sentence (1) in context (2b).

One of the factors that affected the decision of the number of main items is the effect of lexical items. Schütze and Sprouse (2014) claims that the experiment should idealy include 8 or more items to minimize the effect of particular lexical items. In fact, as far as the power of the experiment is concerned, the experiment should have as many items as possible. However, the experiment will involve only 16 items because long experiment may decrese subjects' performance due to tiredness or boredom. Since the experiment will use the Latin square design, each participant see only one condition from each item, meaning that each of them sees only 16 main sentences. However, the experiment will involve twice as many filler sentences as the main sentences following Cowart (1997). Thus, each participant will see 48 sentences in total. Due to the complexity of data to be presented to the participants, the experiment is expected to take around 30 minutes includuing the time for instructions, and brief questionnaires.

Among the 32 filler items, I treat 6 sentences as anchor sentences in light of using a 7-point Likert scale method (1 is least natural and 7 is the most natural). While acceptability judgment seems to be a straightforward task in general, it is necessary to ensure that all the participants use the scale in the same way. Thus, at the beginning of the experiment, 6 anchor sentences will be provided; two of them are clearly unnatural sentences (acceptability = 1 or 2), other two

[1] The number of participants will be discussed in more detail in Section 4.4.

sentences are the ones with controversial acceptability (acceptability = 3, 4, or 5), and the other two sentences are clearly natural sentences (acceptability = 6 or 7). These anchor sentences as well as other filler sentences will encourage participants to provide responses with all the acceptability ratings from 1 to 7, in turn prevents the scale bias such as skew and compression. They will also prevent participants getting aware of the primary interest of this experiment, namely the presence of a particular contextual effect on the availability of the accusative case in Japanese.

3.3 Procedures

The experiment will take place online. After participants answer the questionnaire about their language background and familiarity with linguistics, they will first learn what this experiment means by saying that a sentence is natural or unnatural. Specifically, the instruction will make sure that participants factor out the prescriptive grammar rules, likelihood of the the sentence being uttered in the actual world, plausibility of the content of the sentence and meaningfulness (as opposed to grammaticality/felicity). Following Schütze and Sprouse (2014), the experiment will instruct these points by providing the Japanese counterparts of the following sentences.

(6) In determining if a sentence sounds natural or unnatural, you can imagine that you are talking with your friend and consider whether the sentence would make them sound like a native speaker of Japanese. The experiment is not concerned with whether the sentence is a "good Japanese" as writing, whether the sentence is the best utterance to convey the speaker's idea, how often the sentence is said in normal daily speech, or whether the context provided for the sentence is likely to happen in the actual world. You should also ignore the Japanese grammar you learned at school or any rules you have heard of (e.g., *ra-nuki*speech). The experiment is interested in whether the sentence *could* be said by native speakers. But you should assume that the sentence does *not* involve any production error.

After (6) is presented, participants will be asked to read each sentence and its context carefully, and rate the acceptability of the sentence on a scale between 1 (completely unnatural) and 7 (completely natural).

Except for the anchor sentences, which will always be presented to the participants in the same order, all the other items will be presented to the participant in four different orders in the same line with Sprouse and Almeida (2012): original order, reversed order, transposition of the first and second half items, and reversed order of the transposed order. The experiment employs this counterbalanced design to minimize the order effects mentioned above (i.e., effects of tiredness, boredom, losing intuitions).

4 Power analysis

This section explains how I estimated the number of participants required for this experiment. Given that there seems to be no data frame I can use to carry out simulations, Section 4.1 first creates a data frame. The data frame will be used to fit a model for simulations, but in determining the values of variables in the model, I will refer to the acceptability judgment data from Linzen and Oseki's (2018) experiment about a Japanese phenomenon, which seems to be

the most relevant to the phenomenon to be examined. So Section 4.2 examines Linzen and Oseki's data, and fits a linear mixed-effects model. Based on the examinations, Section 4.3 fit a model to be used for simulations. Finally, Section 4.4 carry out simulations using the `simr` package in R, and discusses the results of simulations.

4.1 Design of a data frame

To fit a mixed-effects model for simulations, if one does not have data related to their experiment, what needs to be done first is to determine what kind of covariates are necessary and create a data frame on which the model will be fitted later. In the current study, I decided to involve four covariates: `item`, `subject`, `condition`, and `order`, which is about the order of the sentence to be presented to the participant. First, since the experiment examines the differnece in acceptability between sentences in two kinds of contexts, the `condition` contains 2 levels. I call those two levels `0` and `1`; they correspond to the copular sentences that can and cannot involve the predicate accusative case, respectively. As for the other covariates, I let them contain 8 levels. This is because the acceptability judgment experiment should involve at least 8 items in general (see Section 3.2) and the number of levels of each covariate can be increased after the first "basic" data frame is created. In other words, I set the number of levels of `item` for the minimun required number of items, and let `subject` and `order` also involve 8 levels so that the number of each covariate can be gragually increased based on the simulation results. For `item` and `subject`, each level is assigned one of the numbers `1-7`. For `order`, each level is assigned n such that n $\in$ {`0`, ..., `7`} and n is interpreted as the order n+1; for instance, when the value of `order` is `0`, it means that a sentence was presented to the subject as the first sentence. By setting the minimum value of `order` as `0`, it will be easier to interpret the effect size of `condition` in the context of `order` = `0`.

After deciding the number of levels of each covariate, we apply the `data.frame` function to the covariates in the way that participants see only one condition in each item and the sentences are presented to them in the different order.[2] The first 16 rows of the data frame created in this way look as in Figure 1 below.

[2]The data frame does factor into the filler and anchor items, and the order of the data is not exactly the same as mentioned in Section 1, due to some technical reasons. But the power analysis for the current experiment needs to "guess" various values as discussed below, so the differences in those aspects of the design would not significantly affect the estimation of the power.

item	condition	subject	order
1	0	1	0
1	0	2	1
1	0	3	2
1	0	4	3
1	1	5	4
1	1	6	5
1	1	7	6
1	1	8	7
2	0	8	0
2	0	7	7
2	0	6	6
2	0	5	5
2	1	4	4
2	1	3	3
2	1	2	2
2	1	1	1

Figure 1: The first 16 rows of the data frame created for simulations.

4.2 Linzen and Oseki (2018)

Linzen and Oseki (2018) examine the acceptability judgments of several Japanese phenomena in the literature, which appear to be questionable by the authors. One of them is a phenomenon about the relation between morphological cases and sentence meanings in Japanese, as exemplified in (7).

(7) a. Taro-wa migime-dake-**o** tumur-e-ru.
Taro-Top right.eye-only-Acc close-can-Prs
'Taro can wink his right eye.'

b. *Taro-wa migime-dake-**ga** tumur-e-ru.
Taro-Top right.eye-only-Nom close-can-Prs
'Taro can wink his right eye.'

(Linzen and Oseki 2018, 8)

Tada (1992) observes that only the sentence with accusative case in (7a) is grammatical with the intended meaning. Linzen and Oseki (2018) find the existence of the contrast between (7a-b) questionable, but the result of their expriment shows that there is an significant difference in acceptability between those two sentences. They report that the difference in mean acceptability ratings between (7a-b) is 1.19 on the scale of 7 point acceptability judgement.

It is worth mentioning that the sequence of words in (7b) is acceptable; the sentence is not acceptable with the "wink reading", but it is acceptable with the meaning that it is only the right eye that Taro can close. So the phenomenon exemplified in (7) is similar to the the phenomenon to be examined in the current experiment in this respect, as well as the fact that both phenomena are related to case and semantics.

In analyzing the acceptability judgment collected in the Likert scale method, I use z-score transformation to minimize the effect of scale compression and scale skew (e.g., Schütze and Sprouse 2014). After obtaining the z-score transformed values of the acceptability judgments, I fitted a linear mixed-effects model. First, the fixed effects in the model is summarized in the table below.

Coefficient	$\hat{\beta}$	SE($\hat{\beta}$)	t	p
(Intercept)	0.72	0.08	9.33	8.80e-15
case	-0.57	0.1	-5.7	1.59e-07
$R^2 = 0.28$, Residual SE = 0.64 (df= 174), n= 178				

The intercept refers to the z-score transformed values of the acceptability judgments of sentence (7a). The variable `case` indicates the difference in the z-score transformed values between (7a) and (7b). On the assumption that α level is 0.05, the table shows that `case` has a significant effect on the z-score transformed values of the acceptability judgments. ($p < 0.001$).

The following table summarizes values of random effects and residuals.

Groups	Name	Variance	Std.Dev.
subject	(Intercept)	0.1159	0.3405
subject.1	case	0.0721	0.2685
Residual		0.412	0.6419

The table does not involve any item-by random effect because Linzen and Oseki (2018) uses only the set of sentences in (7) as an item to examine the effect of morphological case on the meaning in question.

Based on the values of variables in the above tables, the next section fits a model to be used for simulations.

4.3 Fitting a model

This section will fit the linear mixed-effects model in (8).

(8) y ∼ condition + order + (1 + condition + order || subject) + (1 + condition + order || item) + ϵ

In what follows, the fixed effects in the model in (8) will be first estimated (Section 4.3.1), and then random effects and residual will be estimated (Section 4.3.2).

4.3.1 Fixed effects

This section discusses the three fixed effects: intercept, `condition`, and `order`. I explain why the model in (8) should involve those fixed effects but not the interaction `condition`×`order`, and how I determined the effect size of each variable. First, I set the fixed intercept as 0.6, which is slightly smaller than the fixed intercept in the model for Linzen and Oseki's data. What this means is that when `order` = 0, Acc.yes sentences are predicted to be slightly less grammatical than sentence (7a) on average. But it should be noted that the value of the fixed

intercept does not significanlty affect the power this paper is interested in, as illustrated in in Section 4.4.

Next, I turn to the `condition`. This variable is concerned with the contextual effect on the availability of the accusative case, which the current experiment is primarily interested in. Thus, the variable must be included in the model so that the power for detecting the effect can be calculated later. As for its fixed effect size, I assume its value to be -0.5, as opposed to the `case` having the value of -0.57. The slightly larger value of `condition` reflects the prediction that the difference in grammaticality between Acc.yes and Acc.no sentences is smaller than the difference between (7a) and (7b). This effect size will be used in simulations as a baseline, I will investigate the power for detecting a range of different effect sizes.

Next, I turn to the variable `order`, which is about the order of sentences to be presented to participants. Through my consultation experience, the more examples people consider, the higher ratings they tend to give to sentences uttered in both conditions. Two factors that would contribute to this tendency are:

(9) a. Japanese copular sentences with a predicate case are not used in writing very frequently.

b. People say those sentences, but their counterparts without a predicate accusative case is much more common.

Because of these factors, some people may first find the main sentences in the experiment unnatural regardless of their utterance context. But they may realize through the experiment that those sentences are indeed used in normal daily life. This is why I assume a positive slope for `order`. Specifically, I assume its effect size to be 0.03. First, since the effect of orders on sentence acceptability varies in each experiment in size (and direction),[3] I made an assumption that on average the acceptability increases by 0.2 point from `order` = 0 to `order` = 7. Then, given that the effect size of `order` indicates the predicted change in acceptability when the order is changed from 0 to 1, the ratio of 0.2 to 8 (number of levels of `order`) is approximately equivalent to the ratio of 0.03 to 1. Thus, I assume the effect size for `order` to be 0.03. It should be noted that the fixed effect of `order` will not affect the power of detecting `condition` very much.

Lastly, I explain why I decided not to involve the interaction `condition`×`order` in the model in (8). The decision was made by considering the relation between the two factors in (9) and two levels in `condition`. Essentially, there seems to be no solid reason to assume that there is a significant difference in the effect of the two factors in (9) on the acceptability between Acc.no and Acc.yes sentences. Admittedly, if many participants give a 7 point to Acc.yes sentences from the beginning of the experiment, the effect of the factors in (9) can only contribute to the increase of the acceptability of Acc.no sentences. However, it is not likely that even Acc.yes sentences often receive a 7 point due to presence of anchor and filler sentences, for some of which the experiment will use noncontroversially natural sentences. Therefore, I decided not to involve the the interaction `condition`×`order` in the model in (8).

The above discussion about the fixed effects is summarized in the following table.[4]

[3]This observation is based on fitting some models for the English acceptability judgment data provided by Sprouse and Almeida (2017).

[4]The table was created after setting up the values for the random effects and residuals, which will be discussed in the following subsection.

Coefficient	$\hat{\beta}$	SE($\hat{\beta}$)	t
(Intercept)	0.6	0.29253	2.051
condition	-0.5	0.23203	-2.155
order	0.03	0.04264	0.704

R^2 = 0.8, Residual SE = (df= 0.7), n= 64

4.3.2 Random effects

This section discusses the five random effects represented in (8): subject-by and item-by random intercepts, random slope for `condition`, and random slopes for `order`.[5] I explain that it is sensical to involve those random effects, and how I determined values for each random effect. This section also discusses the residual standard deviation of the model in (8).

First, it is sensical to involve random intercepts for both `subject` and `item` because it is likely that the acceptability of Acc.yes sentences presented in `ordrer` = 0 varies among subjects and items. While the presence of filler and anchor ietms would prevent a significant individual variability, I predict that acceptability of Acc.yes sentences will vary more than that of sentence (7a). Given that the subject-by intercept in model.LO has the standard deviation of 0.34, I set the standard deviation of the subject-by intercept in the model in (8) as 0.45. This number indicates that 95% of the subject-by intercepts lies within the 2 standard deviation interval -0.3 (0.6 - 0.45*2) and 1.5 (0.6 + 0.45*2)

I set the item-by intercept as 0.45 as well. The reason is that the sequence of words in different Acc.yes sentences seems to have different frequency of being used in daily life, and more frequent sentences are predicted to receive higher ratings. For example, consider (1) again, which is repeated below as (10).

(10) kyoo-wa onigiri-**o** mit-tu dayo
today-Topic rice.ball-Accusative 3-Classifier copula
'Today is three rice balls.'

It is assumed that the Acc.yes sentence in (10) (or the structure *today is <food name>* in more general) is more common than other Acc.yes sentences such as (11).

(11) kongetu-wa A-gumi-no seeto-**o** san-nin desu
this.month-Topic A-class-Genitive student-Accusative 3-Classifier copula
'This month is three students in class A.'

It is predicted that sentence (11) tends to receive a lower rating than sentence (10) due to the lower frequency of being used in daily life. Therefore, it is sensical to involve the item-by random intercept, and I assume its standard deviation to be identical to that of subject-by random itercept.

Next, I turn to the random slope for each variable. First, it should be noted that both `condition` and `order` are observation-level variables. So it is possible that there is an individual/item variability for the effect of those variables . Having said this, I first discuss the subject-by random slope for `condition`. As the implementation of the current experiment is

[5]It is claimed that "maximal random effect structure" (Barr et al. 2013) often causes convergence problem (e.g., Sonderegger et al. 2018) and there is a high possibility that the analysis of the experimental result will not take random effect correlations into account. Thus, I do not include correlations in the model in (8).

motivated by some individual variability in the acceptability of the main items, the subject-by random slope for `condition` should be included in the model. I wiill set its standard deviation to be 0.3 in light of the standard deviation of the random slope for `case` which is 0.27.

Next, I turn to the item-by random slope for `condition`. It is natural to assume that different items have different differences in acceptability between Acc.yes and Acc.no sentences; for instance, it can be imagined that items with a low frequency of the sequence of words have a larger difference in acceptability between Acc.yes and Acc.no sentences than items with a high frequency of the sequence of words. So it is sensical to involve the item-by random slope for `condition`. As for its standard deviation value, it is difficult to estimate it. So I simply assume it to be idential to the subject-by random slope for `condition`; that is, the stanard deviation is 0.3.

Next, I turn to the random slopes for `order`. First, it seems to be sensical to include both the subject-by and item-by random slopes for the variable. This is because it is reasonable to assume that the effect of order on the acceptability of Acc.yes sentneces should differ among subjects and items. In other words, some people may assign higher ratings to the Acc.yes sentences presented later due to the factors in (9) than others, and some Acc.yes sentences may have larger difference in acceptability depending on whether they are judged toward the beginning or end of the expriment; for example, it can be imagined that items with a frequent sequence of words would be affected by the order less significantly since they wouldn't be affected by the factors (9) very much. As for the exact values of the subject-by and item-by random slopes for `order`, I assume the standard deviations to be 0.03. This number is one standard deviation of the effect of changing `order` on This number indicates that when the acceptability x of an Acc.yes sentence with `order` = n is compared with the acceptability y of an Acc.yes sentence with `order` = n+1, 68% of cases are such that y is within the iterval between x (i.e., x + 0.03 (fixed effect of `order`) - 0.03 (1 sd of the randam slope for `order`)) and x + 0.06 (i.e.,x + 0.03 (fixed effect of `order`) + 0.03 (1 sd of the randam slope for `order`)) For instance, we can compare the acceptability of Acc.yes sentence with `order` = 0 and that of Acc.yes sentence with `order` = 7. The former acceptability is the fixed intercept 0.6. 68% of cases are such that the latter acceptability is within the interval between 0.6 (i.e., 0.6 (fixed intercept) + 0.03 (fixed `order`)*7(distance between `order`=0 and `order`=7) -0.03 (1sd of random `order`)*7(distance between `order`=0 and `order`=7)) and 1.02 (i.e., 0.6 (fixed intercept) + 0.03 (fixed `order`)*7(distance between `order`=0 and `order`=7) + 0.03 (1sd of random `order`)*7(distance between `order`=0 and `order`=7)). Notice that the difference between 0.6 an 1.02 is 0.42, which seems to be a large value/effect of `order` given that the fixed `condition` is -0.5. So the value of random slopes for `order` 0.03 would be conservative value to avoid Type 2 error. Also, the random slopes for `order` do not significantly affect the power this paper is interested in. So the 0.03 should be a good enough estimation.

Finally, I simply assume the residual standard deviation to be 0.7 which is slightly larger than the residual standard deviatio in model.LO which is 0.64. I will use 0.7 as a baseline, and carry out simulations with a range of residual standard deviations in the next section.

Now that all the values for the fixed and random effects as well as residuals are determined, the lmer model in (8) can be fitted using the `makeLemer` function from the `simr` package. The random effects and residuals are summarized below.

Groups	Name	Variance	Standard Deviation
subject	(Intercept)	0.2025	0.45
subject.1	condition	0.0900	0.30
subject.2	order	0.0009	0. 03
item	(Intercept)	0.2025	0.45
item.1	condition	0.0900	0.30
item.2	order	0.0009	0.03
Residual		5.2900	0.7

4.4 Simulation

Based on the model created in the previous section, this section calculates the power of detecting the effect of `condition` using the Monte-Carlo simulation methods. To calculate the power, I apply the `powerSim` function from the `simr` package to the model in (8) whose variables' values were discussed in the previous sections. The result of carrying out the simulation 1000 times indicates the power of detecting the effect with $\alpha = 0.05$ is 0,74, which is lower than the power researchers conventionally aim to achieve, namely 0.8.

In order to increase the power, we can increase the sample size. Technically we can increase the number of either participants or items, but I increase the number of participants here. The reason is that it is not ideal to increase the number of items to keep the experiment short, as mentioned in Section 3.2. After increasing the number of participants from 8 to 24 using the `extend` function, we can carry out simulations for a range of different participant numbers using the `powerCurve` function. The result of calculating the power of the model with 8, 12, 16, 20, and 24 participants is shown in Figure 2.

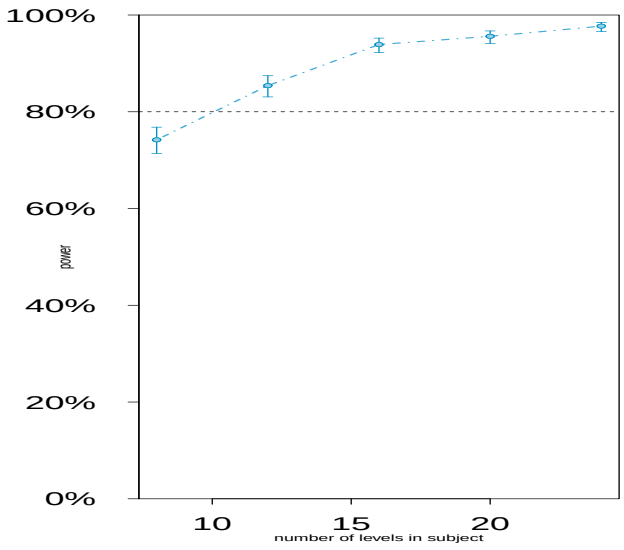

Figure 2: Power curve based on 1000 times simulation of the original model for a range of participant numbers:8, 12, 16, 20, and 24. The y and x axes indicate the power and number of participants. The dashed line indicates power of 0.8.

The figure indicates that when the experiment recruits 12 participants, the original model has power of 0.85 (95% confidence interval : 0.83, 0.88) in detecting the effect of `condition`. Therefore, the experiment should recruits at the very least 12 participants.

The power illustrated in Figure 2 is an estimation that is accurate only if all the values of the parameters in the original model (i.e., fixed effects, random effects, and residuals) are estimated accurately or conservatively. However, it is very possible that the simulation based on the original model overestimates the power. Thus, the rest of this section investigates how many participants would have been required if some variables had been assigned anticonservative values. As mentioned above, while some variables have a large effect on the power, others do not. I first discuss nonsignificant variables. Table 1 shows nonsignificant fixed variables.

	original model	Intercept: 0.6 $\rightarrow$ 0.3	order: 0.03 $\rightarrow$ 0.3
Power	74.20% (71.37, 76.89)	74.40% (71.58, 77.08)	74.60% (71.78, 77.27)

Table 1: Powers and 95% confidence intervals of original model, models where the fixed intercept is lowered to 0.3, and model where the fixed `order` is raised to 0.3, when the number of simulations is 1000.

The fixed intercept and `order` do not affect the power significantly; the power does not change significantly even though the intercept in the original model is changed to its half value or the original `order` is multipled by ten.

Next, I turn to nonsignificant random variables, which are shown in Table 2.

	original model	Intercept: 0.45 $\rightarrow$ 0.9	order: 0.03 $\rightarrow$ 0.3
Power	74.20% (71.37, 76.89)	73.00% (70.13, 75.73)	70.90% (67.98, 73.70)

Table 2: Powers and 95% confidence intervals of original model, models where the subject-by and item-by random intercepts is raised to 0.9, and model where the subject-by and item-by random `order` is raised to 0.3, when the number of simulations is 1000.

Table 2 shows the cases where the intercept in the original model is doubled and the original `order` is multipled by ten. While the powers in those two case slightly differ from the power of original model, the difference is subtle given that the values of *two* variables (e.g., subject-by and item-by intercepts) are largely changed in both cases.

Unlike the variables in Table 1 and Table 2, the other variables have a significant effect on the power this paper is interested in. First, if one changes the fixed `condition` from -0.5 to -0.4, the power of the original model (i.e., 74.20% (71.37, 76.89)) becomes 57.70% (54.57, 60.79). As Figure 3 shows below, to reach the power of 0.8, the experiment requires 20 participants (84.80% (82.42, 86.97)), as opposed to the original model requiring 12 participants.

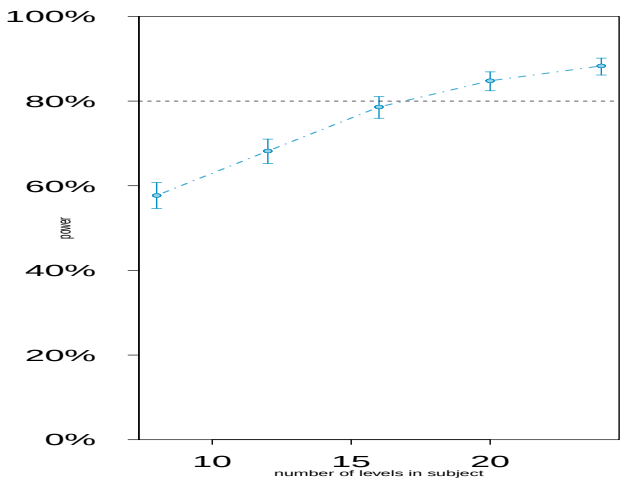

Figure 3: Power curve based on 1000 times simulations of the model with the fixed `condition` = -0.4 for a range of participant numbers:8, 12, 16, 20, and 24.

Next, we consider the cases where we change the subject-by and item-by random slopes for `condition` from 0.3 to 0.4.

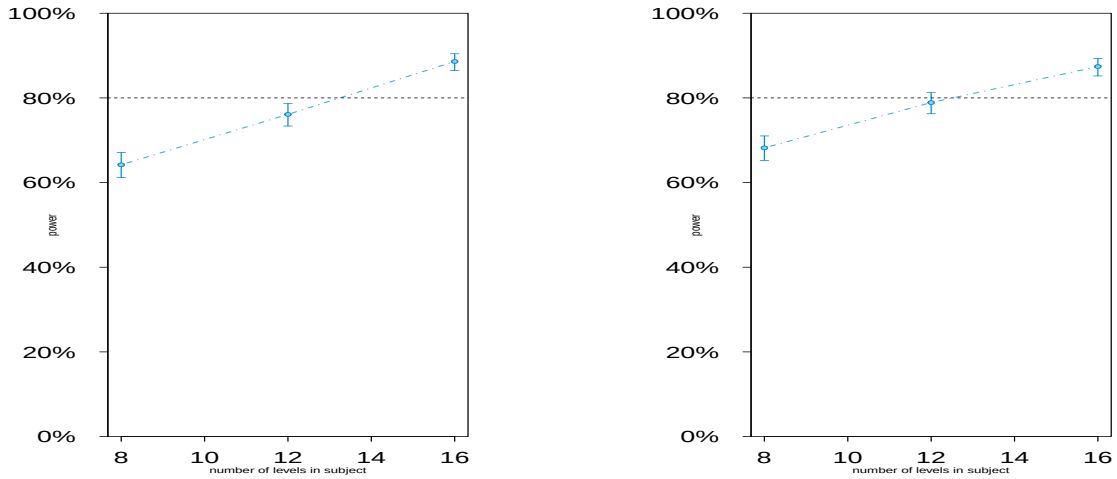

Figure 4: Power curves based on 1000 times simulations of the model with the subject-by (left) or item-by random slope for `condition` being 0.4. Both plots show powers when the number of participants 8, 12, and 16.

The right and left plots in Figure 4 show that when the experiment recruites 16 participants, it can detect the effect with the powers of 0.88 (86.47, 90.50) and 0.87 (85.18, 89.39), respectively. Whereas the difference in 0.1 of the fixed and `condition` values would not simply be comparable, the effect of the random `condition` on the power seems to be smaller than the effect of the fixed `condition`; remember that the model with the fixed `condition` = -0.4 requires 20 participants.

Next, we consider the effect of residuals on the power. In Section 4.3.2, the residual value was set as 0.7. Here, I examine the power of the model when the residula is 1.13 for the

following reason. First, Lane et al. (2016) claim that the residual standard deviation is usually larger than the random effect variances. In the original model, the random variable with the largest value is the subject-by/item-by intercept. In fact, the residul already has a larger value (0.7) than the intercept (0.45). But there is a question of to what extent the residual can be larger than a random effect with the largest value in general. This question is worth addressing because larger residual values lead to lower power, and thus I should examine powers with a larger residual power. So I referred to Sonderegger et al. (2018) and Winter (2019). I checked all the linear mixed models in those literature, and examined how many times the residual is larger than the largest random effect within the same model at largest. It turns out that the residual is at largest 2.5 times as large as the largest random effects. While this observation is based on a small sample size and it is possible that the difference between the residual and the largest random effects could be larger, this is how I decided to set the residual value as 1.13 ($\approx$ 0.45$\times$2.5). Consider Figure 5 for the power curves of the model with residul =1.13.

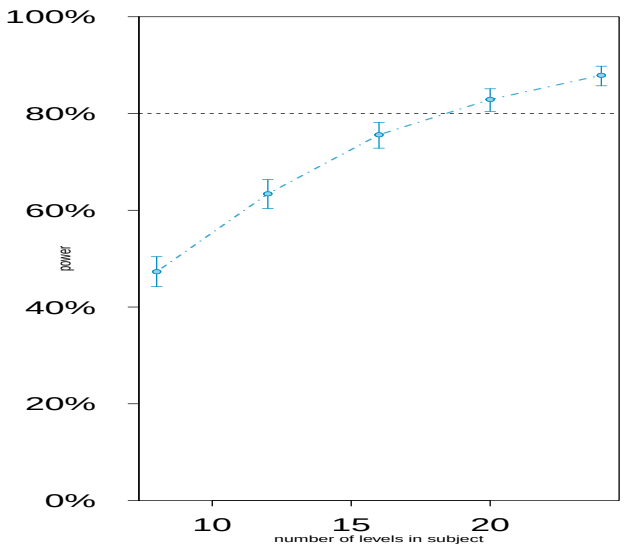

Figure 5: Power curve based on 1000 times simulation of the model with residul =1.13 for a range of participant numbers:8, 12, 16, 20, and 24.

Figure 5 shows that when the experiment recruites 20 participants, it can detect the effect with the powers of 0.83 (80.42, 85.18).

I have so far revised the value of one variable in the original model and exmined the power. But it is possible that the original model involves more than one variable with an anticonservative value. So I finally examine the power of the model where the fixed `condition` = -0.4, subject-by and item-by random slopes for `condition` = 0.4, and residual = 1.13. Consider Figure 6.

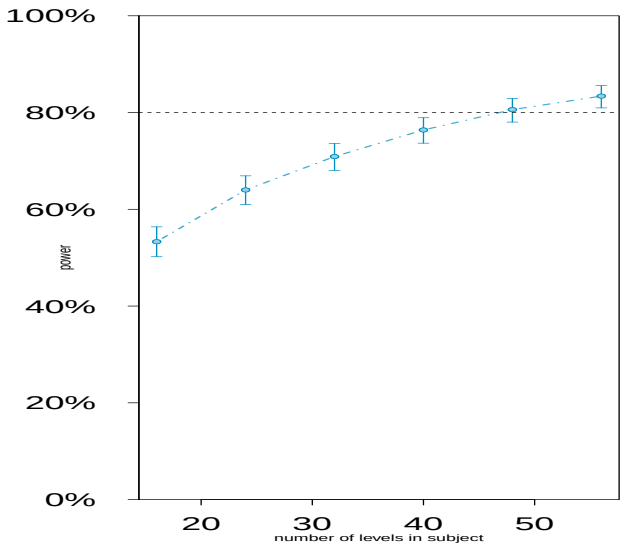

Figure 6: Power curve based on 1000 times simulation of the model where the fixed `condition` = -0.4, subject-by and item-by random slopes for `condition` = 0.4, and residual = 1.13.

Figure 6 shows that when the experiment recruites 56 participants, it can detect the effect with the powers of 0.83 (80.95, 85.66).

The above power calculations show that the number of participantes required to detect the effect of `condition` with the 80% power varies significantly depending on the values of some variables. While the original model requires only 12 participants, the largest number of participants required by a model is 56. In fact, the experiment may actually require more than 56 participants. However, it is also reasonable to respect the accuracy of the original model to some extent given that the values of the variables in the model was estimated based on the analysis of model.LO. Therefore, it would be reasonable to assume 56 as the required number to detect the effect of `condition`.

5 Conclusion

This paper discussed the design of an acceptability judgment experiment. Section 2 first briefly explained the phenomenon to be examined in the experiment; that is whether the acceptability of the predicate accusative case in Japanese copular construction depends on the context where the sentence happens. Section 3 then discussed the experiment's participants, materials, and procedures. In designing the experiment, one of the primary questions concerns how many participants the experiment should recruit. To estimate the answer for this question, Section 4.4 analyzed Linzen and Oseki's (2018) experimental data about a similar Japanese phenomenon, carried out simulations, and concluded that the experiment aims to recruit around 56 participants. While this estimation consists of many "informed guesses", the simulation is definitely of use to prevent estimating too small sample size for the experiment.

References

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

Cowart, W. (1997). *Experimental syntax*. Sage.

Dabrowska, E. (2010). Naive v. expert intuitions: An empirical study of acceptability judgments. *The linguistic review 27*(1), 1–23.

Edelman, S. and M. H. Christiansen (2003). How seriously should we take minimalist syntax? a comment on lasnik. *Trends in Cognitive Science 7*(2), 60–61.

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

Gordon, P. C. and R. Hendrick (1997). Intuitive knowledge of linguistic co-reference. *Cognition 62*(3), 325–370.

Grewendorf, G. (2007). Empirical evidence and theoretical reasoning in generative grammar. *Theoretical Linguistics 33*(3), 369–380.

Harada, M. (2018). Contextual effects on case in japanese copular constructions. In *Proceedings of the 12th Generative Linguistics in the Old World and the 21st Seoul International Conference on Generative Grammar*, pp. 447–456.

Lane, S., E. Hennes, and T. West (2016). *I've Got the Power: How Anyone Can Do a Power Analysis of Any Type of Study Using Simulation*.

Linzen, T. and Y. Oseki (2018). The reliability of acceptability judgments across languages. *Glossa: a journal of general linguistics 3*(1).

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

Schütze, C. T. and J. Sprouse (2014). Judgment data. *Research methods in linguistics 27*.

Sonderegger, M., M. Wagner, and F. Torreira (2018). Quantitative methods for linguistic data. version 1.0.

Spencer, N. J. (1973). Differences between linguists and nonlinguists in intuitions of grammaticality-acceptability. *Journal of psycholinguistic research 2*(2), 83–98.

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

Sprouse, J. and D. Almeida (2017). Design sensitivity and statistical power in acceptability judgment experiments. *Glossa 2*(1), 1.

Tada, H. (1992). Nominative objects in japanese. *Journal of Japanese Linguistics 14*(1), 91–108.

Wasow, T. and J. Arnold (2005). Intuitions in linguistic argumentation. *Lingua 115*(11), 1481–1496.

Winter, B. (2019). *Statistics for linguists: An introduction using R*. Routledge.