A META-ANALYSIS OF RELATIVE CLAUSE PROCESSING IN MANDARIN CHINESE USING BIAS MODELLING

DISSERTATION

Submitted in Partial Fulfillment of the Requirements for

the Degree MSc in Statistics at the

School of Mathematics and Statistics of the University of Sheffield

By

Shravan Vasishth

* * * * *

The University of Sheffield
September 2015

Dissertation Advisor:

Professor Jeremy Oakley

Abstract of: A Meta-analysis of Relative Clause Processing in Mandarin Chinese using Bias Modelling

Author: Shravan Vasishth

Date: September 2015

The reading difficulty associated with Chinese relative clauses presents an important empirical problem for psycholinguistic research on sentence comprehension processes. Some studies show that object relatives are easier to process than subject relatives, while others show the opposite pattern. If Chinese has an object relative advantage, this has important implications for theories of reading comprehension. In order to clarify the facts about Chinese, we carried out a Bayesian random-effects meta-analysis using 15 published studies; this analysis showed that the posterior probability of a subject relative advantage is approximately 0.77 (mean 16, 95% credible intervals -29 and 61 ms). Because the studies had significant biases, it is possible that they may have confounded the results. Bias modelling is a potentially important tool in such situations because it uses expert opinion to incorporate the biases in the model. As a proof of concept, we first identified biases in five of the fifteen studies, and elicited priors on these using the SHELF framework. Then we fitted a random-effects meta-analysis, including priors on biases. This analysis showed a stronger posterior probability (0.96) of a subject relative advantage compared to the standard random-effects meta-analysis (mean 33, credible intervals -4 and 71).

ACKNOWLEDGMENTS

I'm very grateful to Professor Jeremy Oakley for his timely and extremely useful advice during the preparation of this dissertation. The teaching staff of the School of Mathematics and Statistics also deserve thanks for their high quality lecture notes and teaching in the MSc Statistics programme. Finally, I am very grateful to my wife, Andrea Vasishth, for making it possible for me to complete this degree.

Table of Contents

CHAPTER 1

THE ISSUE: PROCESSING CONSTRAINTS ON CHINESE RELATIVE CLAUSES

1.1 Introduction

Psycholinguistics, a subfield of linguistics, focuses on developing theories of language comprehension and production processes, at the word, sentence, and discourse level. Within psycholinguistics, sentence comprehension research is concerned with syntactic and semantic processes unfolding in online language comprehension, both in the written and spoken modality.

Several computationally implemented models of sentence comprehension exist. These models make quantitative predictions about moment-by-moment processing difficulty when native speakers read sentences. There is a broad consensus in the field that both probabilistic knowledge of language (Levy, 2008) and working memory constraints (Lewis and Vasishth, 2005) affect the speed and accuracy of word-by-word comprehension processes; in many cases, it is also clear that fairly subtle linguistic constraints can be deployed by the human language comprehension system (henceforth, the parser) to build structure (Stowe, 1986).

Several experimental methods are standardly used to compare the predictions of these models with data from human subjects engaged in language comprehension. A commonly used method is self-paced reading (Just et al.,

1982). Here, the subject is seated in front of a computer screen. Each trial begins with a series of dashes on the screen. When the subject presses the space bar on the keyboard, the first word appears. When the space bar is pressed again, the first word is replaced again with dashes and the next word is uncovered; in this way, the subject reads the sentence word by word or phrase by phrase, and the experiment software is able to record the time in milliseconds spent on each word or phrase. An example of how the screen unfolds is shown in Figure 1.1.

```
--- ----- ----- ---- --- ---- ----.
The ----- ----- ---- --- ---- ----.
--- horse ----- ---- --- ---- ----.
--- ----- raced ---- --- ---- ----.
--- ----- ----- past --- ---- ----.
--- ----- ----- ---- the ---- ----.
--- ----- ----- ---- --- barn ----.
--- ----- ----- ---- --- ---- fell.
```

Figure 1.1: A schematic illustration of the self-paced reading method.

In a self-paced reading (SPR) study, reading times are recorded from multiple subjects reading sentences that have a theoretically interesting experimental manipulation (an example is discussed below), leading to repeated-measures data. Such experiments have the danger that the subject may stop paying attention to the sentences; to forestall this, subjects are usually asked comprehension questions after each sentence, and comprehension accuracies are informally used to ensure that subjects were attending to the task. Thus, comprehension accuracy can be used as an approximate guide to how deeply

the subject is processing the material (this assumes that answering the questions requires a complete understanding of the sentence.)

SPR has the great advantage of simplicity: the subject reads every word/phrase in sequence and the time spent on each word/phrase is taken as an estimate of the time taken to complete syntactic and semantic processing. This usually leads to a completely balanced data-set with no missing values. SPR is also very convenient for investigating less well-studied languages because one can travel to the field and conduct experiments there, without anything more sophisticated than a laptop.

Another commonly used method is eyetracking while reading. Here, the subject is seated in front of a computer and their eye movements are recorded while they read a sentence on the screen. This method has the advantage that, compared to SPR, a more natural record of the reading process is obtained. Apart from requiring more technical knowledge than SPR, the principal disadvantage is the increased complexity of analysis: readers skip short, high frequency words, and make leftward eye movements (regressions) to revisit previously read words. Due to the fact that most trackers are not portable, this method is usually used only in laboratory settings and not in the field (although portable trackers do exist).

Thus, SPR and eyetracking both deliver reading times in milliseconds for each word or region in a sentence. These are assumed to reflect comprehension difficulty, and can therefore be compared to the predictions of computational models of sentence comprehension. We restrict attention in this dissertation to such reading studies.

1.2 Chinese relative clauses

Relative clauses (RCs) have received a lot of attention when evaluating predictions of sentence comprehension theories. This is because RCs have certain properties that are ideal for comparing the predictions of alternative theories.

Consider a sentence such as *The man was reading a book*. This sentence contains a subject (*man*) and an object (*book*), and the meaning of sentence (roughly, who did what) arises from the link or dependency between the subject and object with the verb phrase *was reading*. An RC involves modification of one of these nouns by a so-called relative clause. Consider example 1.

(1) The man that greeted the doorman was reading a book

Here, the proposition in the relative clause, *The man greeted the doorman*, can only be computed fully by the reader once the relative pronoun *that* is associated with *man*. The sentence in example 1 is a subject relative clause (SRC) because the *man* is the subject of the RC.

One can also build an object relative clause (ORC); see example 2.

(2) The man that the doorman greeted was reading a book.

The theoretically interesting issue here is that completing the syntactic dependency between the subject *man* and the RC verb *greeted* has been found to be more difficult to complete in object relatives than subject relatives. One explanation that has been proposed for this difference between ORs and SRs is that the distance between the subject and the RC verb is longer in the

OR compared to the SR (Gibson, 2000; Lewis and Vasishth, 2005). This so-called SR advantage has been widely considered to be a linguistic universal, because in virtually every language of the world, subject relatives are easier to process than object relatives.

This explanation for the SR advantage, which we can call the **dependency distance** account, makes the surprising prediction that in Chinese, object relatives will be easier to process than subject relatives. This is because in Chinese, the RC verb and subject noun distance is longer in SRs than in ORs. This can be seen in the examples shown below. The RC verb-subject noun distance is longer in SRs because relative clauses in Chinese appear before the noun they modify (in English, RCs follow the noun they modify).

(3) a. Subject relative

[**yaoqing** fuhao de] **guanyuan** xinhuaibugui
invite tycoon DE official have bad intentions

'The official who invited the tycoon had bad intentions.'

b. Object relative

[fuhao **yaoqing** de] **guanyuan** xinhuaibugui
tycoon invite DE official have bad intentions

'The official who the tycoon invited has bad intentions.'

The key empirical issue is therefore whether reading time is longer at the head noun (here, *guanyuan*) in ORs compared to SRs. Unfortunately, the literature on Chinese relative clauses has produced quite a mixed picture, which we discuss in the next chapter.

1.3 Objectives in this dissertation

We have two main objectives in this dissertation. First, in psycholinguistics, although literature reviews are published routinely, there is no tradition of doing formal meta-analyses. One goal is to synthesize our current evidence relating to Chinese relative clauses. Second, we are interested in quantifying, through expert judgements, the extent of the biases present in the studies, in order to obtain less biased estimates of the true effect. Beyond these two objectives, a broader goal is to develop a methodology for psycholinguistics that can be used to carry out bias modelling on a larger scale. Bias modelling can greatly help improve our understanding of open questions in psycholinguistics. For example, Engelmann et al. (2015) report a comprehensive review of 69 published studies on sentence comprehension. This review reveals surprisingly large heterogeneity between studies, suggesting a need for bias modelling to uncover more accurate estimates of effects.

CHAPTER 2

A PRELIMINARY EXAMINATION OF THE AVAILABLE DATA

2.1 Introduction

In this chapter, we summarize the data available on Chinese relative clauses, and conduct some exploratory analyses. We show that the effects (on the millisecond scale) seem to belong to two distributions, suggesting that studies may have different sources of bias. We also show that there is some indication of publication bias: more exaggerated effects seem to have been published than would be expected under repeated sampling. Finally, we establish through a Monte Carlo Hypothesis Test that the between-study variance is larger than might be expected under random sampling. All this suggests that, in addition to publication bias (discussed below), the data may have systematic biases that mask the true effect.

2.2 The data on Chinese relative clauses

Table 2.1 summarizes the available data on Chinese relatives. In this dissertation, we only consider studies done on native speaker adults, and studies which aim to compare processing differences in Chinese subject and object relatives.

	study	y	se	nsubj	nitem	qacc	method	location
1	Gibson et al 12	-120	48	37	15	91	SPR	Taiwan
2	Vas. et al 13, E3	-109.40	54.80	40	15	87	SPR	Dalian
3	Lin & Garn. 11, E1	-100.00	**30.00**	48	80	88	SPR	Taiwan
4	Qiao et al 11, E1	-70.00	42.00	32	24		GMaze	USA
5	Lin & Garn. 11, E2	-30.00	44.63	40	80		SPR	Taiwan
6	Qiao et al 11, E2	6.19	19.90	24	30		LMaze	Shanghai
7	Hsiao et al 03	50.00	25.00	35	20	70	SPR	USA
8	Wu et al, 11	50.00	**40.74**	48			SPR	Shanghai
9	Wu 09	50.00	23.00	40			SPR	Shanghai?
10	Jaeg. et al 15, E1	55.62	65.14	49	16	85	SPR	Nanjing
11	Chen et al 08	75.00	35.50	39	23	86	SPR	Beijing
12	Jaeg. et al 15, E2	81.92	36.25	49	32	80	ET	Taiwan
13	Vas. et al 13, E2	82.60	41.20	61	24	82	SPR	Dalian
14	C Lin & Bev. 06	100.00	80.00	48	24		SPR	Taiwan
15	Vas. et al 13, E1	148.50	50.90	60	20	82	SPR	Taiwan

Table 2.1: Summary of the 15 data-sets on relative clauses. The column y refers to the effect size (in milliseconds) of each study (negative values are an object relative advantage); se refers to the estimated standard error in each study; nsubj refers to the number of subjects in each study and nitems to the number of items; qacc refers to average question-response accuracy (averaging over relative clause types); method refers to the experimental method used (SPR is self-paced reading, GMaze and LMaze are the maze tasks described in the main text, and ET refers to reading using eyetracking); and location is the country where the data were collected. Blank spaces represent missing information. The standard errors in bold are imputed by taking the mean of all the standard deviations s and then computing the standard error from them using the formula $s/\sqrt{n}$, where n is the study subject sample size.

Most of these studies are self-paced reading experiments; the only exceptions are those by Qiao et al. (2012), and Jäger et al. (2015). Jäger and colleagues use eyetracking while reading, and Qiao and colleagues use a method similar to self-paced reading called the maze task. In the maze task, subjects press the space bar to see two alternative continuations of a sentence, and have to choose one continuation over the other. For example, if subjects have already read *The man who*, in the next stage of the task, they could be made to choose between words like *the* (which would imply an OR continuation) and *hired* (which would imply an SR continuation). The time taken to choose the correct alternative is taken as the processing time. In a variant of the maze task, also used by Qiao and colleagues in their Experiment 2, the subject has to choose between a word that could continue the sentence, and a non-word that is created by combining two legal Chinese characters which do not together form a word. This variant allows the subject to reject the illegal continuation by making a lexical decision when faced with a choice.

All the studies in Table 2.1 investigate subject vs object relatives having a structure similar to the sentences shown below:

(4) a. Subject relative

[**yaoqing** fuhao de] **guanyuan** xinhuaibugui
invite tycoon DE official have bad intentions

'The official who invited the tycoon had bad intentions.'

b. Object relative

[fuhao **yaoqing** de] **guanyuan** xinhuaibugui
tycoon invite DE official have bad intentions

'The official who the tycoon invited has bad intentions.'

Such relative clauses are also called single embedded relatives, because a single relative clause is embedded within a main clause. In principle, it is also possible to have double embedded relatives. Here, a relative clause is embedded inside the relative clause, which itself is embedded inside a main clause; an example from Sampson (2001) is

(5) Don't you find that sentences that people you know produce are easier to understand?

Such sentences are, strictly speaking, grammatical, because syntactic rules in English and other languages allow relative clause modification of any noun, regardless of how deeply embedded the noun is. However, although they are grammatical, double embeddings in general are very difficult to understand in English and many other languages. The reason for this difficulty probably has to do with limitations of human working memory and exposure (Gibson and Thomas, 1999; Vasishth et al., 2011; Frank et al., 2015).

The first major study on Chinese relative clauses (Hsiao and Gibson, 2003) had single as well as double embeddings in a repeated measutes design; thus, Hsiao and Gibson had a 2×2 factorial design, with Relative Clause type and Embedding as factors. Our meta-analysis focuses only on single embedded relative clauses because, as in English, double embeddings are unusually difficult to understand in Chinese. This means that we ignore results from the double embedding conditions in the Hsiao and Gibson (2003) study. Furthermore, our analysis only looks at reading time at the head noun (the noun modified by the relative clause), because our focus is on

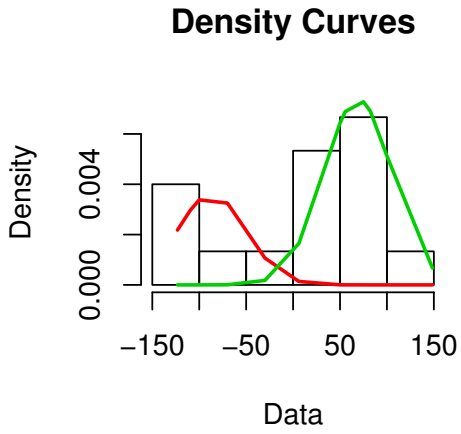
investigating the evidence for a very specific hypothesis, the dependency distance account (see page 5 for an explanation). A more thorough analysis would look at reading times in other regions of the sentence as well, but this beyond the scope of the present dissertation.

2.2.1 Data extraction

We had the original raw data from eight of the fifteen studies. In these cases, the estimate of the effect (the difference between subject and object relatives) was calculated by fitting a linear mixed model (Bates et al., In Press), with raw reading time as a dependent variable; this gave us the estimate of the mean difference and its standard error. In the remaining experiments, for which we did not have the raw data, we estimated the mean either from the figure or from reported tables of means; standard error was estimated either using the reported t- or F-value. For example, if a mean effect of $\bar{x}$ ms is reported in a study, and an absolute t-value of t is reported, then we can compute the estimated standard error by solving for SE in the formula $t = \frac{\bar{x}-0}{SE}$. A reported F-score can be converted to a t-value by using the fact that $t^2 = F$. When neither the t- of F-value was reported, we examined the figure in the paper and estimated the standard error visually by measuring the width of the confidence interval. Half the width would give us approximately 1.96 times the standard error. In two cases (Lin and Garnsey, 2011; Wu et al., 2011), not enough information was provided to derive standard error estimates, and in these cases we took the mean of the standard deviations from the other studies and divided by the square root of

the subject sample sizes from each of these two studies to impute standard error.

2.2.2 The distribution of the effect sizes



parameter	Distribution 1	Distribution 2
lambda	0.33	0.67
mu	-86.67	69.03
sigma	36.29	36.292

Figure 2.1: The mixture distribution of effects across the Chinese relative clause studies investigated.

The effects across the studies have a bimodal distribution, which can be modelled as a mixture of normals; see Figure 2.1. This mixture distribution was estimated using the R library `mixtools`; the function `normalmixEM` in this library uses the standard Expectation Maximization algorithm (McLachlan and Peel, 2000) to estimate the parameters of the mixture distributions. Negative estimates (the column y) in Table 2.1 mean that an OR advantage was found, and positive estimates mean that an SR advantage was found.

2.3 Checking for evidence of publication bias

A bimodal distribution of effects across studies could arise due to systematic differences between studies. We will attempt to model some of the possible sources of bias in a later chapter.

Another possible factor that may cause a bimodal distribution of effects is publication bias: it is likely that only studies that show larger effects (in either direction) tend to get published. A further possible cause for the bimodal distribution is the generally low statistical power of studies done in areas like psychology (Cohen, 1988). We discuss next the consequences of low power on the pattern of published results.

2.3.1 Type S and Type M errors in under-powered studies

Gelman and Carlin (2014) have pointed out that low-powered studies can lead to two kinds of error, which they call Type S (sign) errors and Type M (magnitude) errors. Type S error is defined as the probability that the sign of the effect is incorrect, given that (a) the result is statistically significant, or (b) the result is statistically non-significant, and Type M error is the expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size (conditional on whether the result is significant or not). Gelman and Carlin also call Type M error the exaggeration ratio, which is perhaps more descriptive than "Type M error".

Type S and M errors have the consequence that one can end up with the incorrect sign of the effect, and the magnitude of the effect can be dramatically exaggerated. Since journals prefer to publish only statistically significant effects, with lower p-value preferred over marginal ones, it is likely that

the published literature has quite a few exaggerated effect estimates, with the wrong sign. It is easy to illustrate this point with a simulation. Suppose that a particular study has standard error 46, and sample size 37; this implies that standard deviation is $46 \times \sqrt{37} = 279$. These are representative numbers from psycholinguistic studies, and are based on the Gibson and Wu (2013) study. Suppose also that we know that the true effect is D=15. Then, we can compute Type S and Type M errors for replications of this particular study by repeatedly sampling from the true distribution.

1. Take n repeated samples of size 37 from the distribution $N(15, 279^2)$, computing the mean of the i-th sample d_i each time.

2. Assuming that the standard error is known to be 46, compute the absolute t-value d_i/SE and compute the proportion of cases in the n replications that this value is greater than 2. This is our power, and comes out to approximately 0.05 for this particular example.

3. Type S error given that the effect was statistically significant at $\alpha = 0.05$ is 0.2 and the Type S error given that the result is not significant is 0.39.

4. Type M error or the exaggeration ratio under statistical significance is 7.29 and under non-significance it is 2.27.

Based on experience with previous studies, we estimate that a plausible range of effect sizes for the relative clause issue is 15-30 ms. In published psycholinguistic studies, we see large variability in the reported effect sizes for any given phenomenon. For example, in English relative clause studies,

where the subject relative advantage is uncontroversial, in self-paced reading studies, at the critical region (which is the relative clause verb in English), we see 67 milliseconds (SE approximately 20) (Grodner and Gibson, 2005); 450 ms, 250 ms, 500 ms, and 200 ms (approximate SE 50 ms) in experiments 1-4 respectively of Gordon et al. (2001); 20 ms in King and Just (1991) (their figure 6). In eye-tracking studies reporting first-pass reading time during reading,[1] we see 48 ms (no information provided to derive standard error) in Staub (2010); and 12 ms (no SE provided) in Traxler et al. (2002). The larger effect sizes summarized here are quite atypical for psycholinguistic studies on relative clauses. Thus, pending a more comprehensive review of the literature spanning multiple methods and languages, we tentatively assume an absolute true effect size of approximately $15 - 30$ milliseconds for Chinese.

If we assume that the true effect is at the upper bound of 30 ms, the values of Type S and Type M errors under the above assumptions are somewhat smaller but still substantial:

1. Type S error given that the effect was statistically significant at $\alpha = 0.05$ is 0.05 and the Type S error given that the result is not significant is 0.27.

2. Type M error or the exaggeration ratio under statistical significance is 3.79 and under non-significance it is 1.25.

[1]First-pass reading time simply refers to the total amount of time spent fixating on a word, counting from the moment that the eye transitions to the word from the left and up to the moment that the eye exits the word to the right. First-pass reading time is widely considered to be a useful measure of processing difficulty while reading (Clifton et al., 2007).

The above simulations give us some indication of how bad the Type S and M error situation can be under two different boundary values for the true effect size. These simulations suggest that, given some plausible assumptions about the underlying parameters, the published work is likely to be reporting exaggerated effects. But this is only speculative; how can we evaluate the extent of publication bias given the studies under consideration?

2.3.2 The evidence for publication bias

Funnel plots for identifying publication bias

One simple graphical method is to plot precision (the inverse of the variance) against the observed effects across studies (Duval and Tweedie, 2000). This method also presupposes that there is some independent basis for specifying the true effect size. If there is no publication bias, a so-called funnel plot should be seen: under repeated sampling, low precision estimates should be widely spread out on either side of the true effect, and for higher precision estimates, the funnel should be narrow, centered around the true mean; see Figure 2.2.

In the funnel plots shown below, we simply used the mean of the effect sizes observed across studies as a proxy for the true (unknown) effect size. As discussed above, the absolute value of the mean (approximately 15 ms) is not an unreasonable estimate of the true effect size. The simulated funnel plots were derived using the following method:

1. For each sample size n ranging from the minimum to maximum in our 15 studies, we sampled repeatedly from a normal distribution with mean and standard deviation equal to the grand mean of the effect

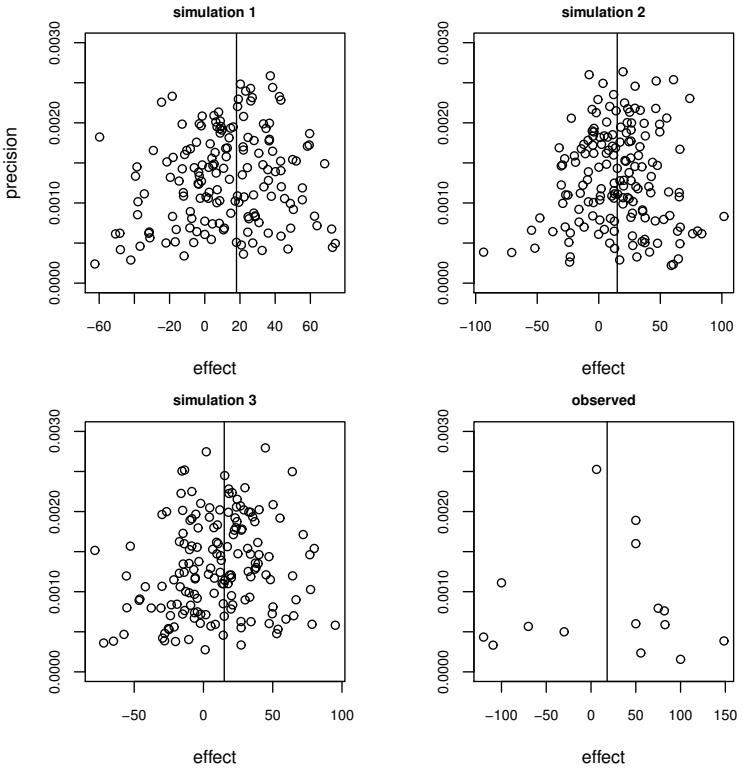


Figure 2.2: Funnel plot for diagnosing publication bias for the present data assuming a true effect size of +15 msec (subject relative advantage). The bottom right plot shows the effect sizes of the data under consideration, and the other plots show funnel plots for simulated data with statistical power comparable to the studies under consideration.

sizes in the 15 studies (approximately 15 ms), and the grand mean of the standard deviation of these studies.

2. After each sample of size n was taken, we computed and stored the sample mean. We also computed precisions for each sample size as $1/(\hat{\sigma}^2/n)$.

3. Finally, we plot precision against the effect size, for each sample size.

This was done three separate times to get a sense of the variability in the shape of the funnel plot; the results are shown in Figure 2.2. Alongside these simulated funnel plots, we also plot the precision of each study against the observed effect size.

Figure 2.2 suggests that there might be a publication bias such that effects with small (and non-significant) studies went unpublished. This is not surprising: in psychology and linguistics, it is difficult to publish non-significant results, and so these would usually go unreported.

Monte Carlo Hypothesis Test to check for exaggerated between-trial variance

One way to check whether exaggerated effects were preferentially published is to test whether the between-study variance is larger than expected under the null hypothesis that the study effects under repeated sampling come from a sampling distribution that is a Normal distribution with mean 15 and standard deviations 42 (the mean of the standard errors observed in the study). We carried out a Monte Carlo Hypothesis Test to test this. The standard method for the Monte Carlo Hypothesis Test is the following:

1. Generate $n-1$ test statistics under H_0.

2. Let $m = n\alpha$.

3. If T_{obs} is one of m largest $\{T_1, \ldots, T_{n-1}, T_{obs}\}$, reject null.

First, we computed between-study variance from the available data (7081), and then ran 9999 simulations, each time generating 15 studies from $Normal(15, 42^2)$. Then, in each iteration, for each set of 15 studies, we computed the variance `var`. This yielded 9999 variances. Letting $m = 1000 \times 0.05 = 500$, if our observed variance is one of the m largest of $\texttt{var}_1, \ldots, \texttt{var}_{9999}$, then we can reject the null hypothesis that the observed between-study variance is typical for studies with these standard errors. Figure 2.3 shows that the observed variance is among the m largest, suggesting that there may be a tendency to publish exaggerated effect sizes (or more accurately, a tendency to not publish small, non-significant effects).

2.4 Summary of preliminary exploration of the data

It is clear from the above discussion using funnel plots and the Monte Carlo Hypothesis Test that, given some plausible assumptions about effect sizes and standard deviation, there is some evidence for publication bias, with a tendency to not publish non-significant effects with a smaller effect size, and the between-study variance is higher than one would expect, suggesting that under-powered studies may be delivering exaggerated effects (Type M error).

A further issue is that biases may exist in each study; these would further skew the estimates from each study. We return to modelling biases later; in

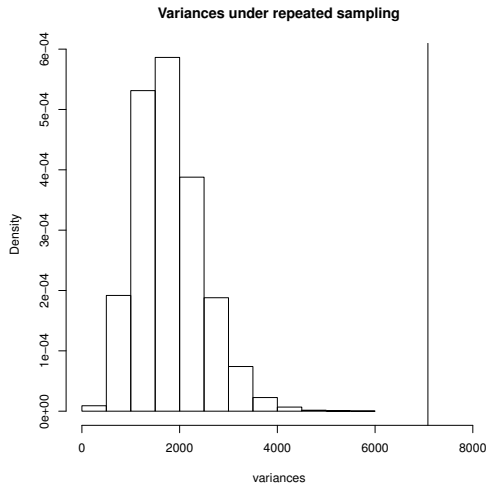


Figure 2.3: Monte Carlo Hypothesis test to check whether observed between-trial variance (vertical line) is typical. The figure shows the distribution of between-study variance of 15 studies under repeated sampling, assuming a generating distribution being a Normal distribution with mean 15 and standard deviation 42 (the mean of the standard errors observed in the 15 studies). We see that the observed variance is much larger than we would expect under the null hypothesis that the studies are sampled from a normal distribution with mean 0 and standard deviation equal to the mean of the standard errors observed in the studies.

the next chapter, we first carry out a random-effects meta-analysis of the available data.

2.5 Concluding remarks

We presented evidence using funnel plots and other methods suggesting that the between-trial variance may be higher than one might expect given the observed effects and their standard deviations. Although there are indications of bias in the data, we will first carry out a standard random-effects meta-analysis in the next chapter.

CHAPTER 3

A RANDOM EFFECTS META-ANALYSIS

3.1 Introduction

In areas like medicine, meta-analysis—using statistical methods to summarize the results of multiple (independent or dependent) studies (Glass, 1976)—has become a well-known method for synthesizing evidence as part of a systematic review of the literature (Higgins and Green, 2008). Systematic reviews in medicine generally aim to bring together all the evidence that meets specific criteria. A primary goal is to use all available evidence to make informed decisions about interventions.

Systematic reviews and meta-analysis allow us to quantitatively take into account the fact that science is cumulative; in the absence of a meta-analysis, only qualitative statements can be made about the state of the art in a particular field. Meta-analysis is, however, not widely used in psychology and linguistics. Instead, the general tendency is to rely on null hypothesis significance testing to evaluate whether a phenomenon has a true effect θ equal to zero or not (see, for example, Engelmann et al. (2015)). The Chinese relative clause issue has also suffered from this problem, with researchers merely noting the disagreements between studies without trying to collate the quantitative evidence. Although potential sources of bias are often recognized in literature reviews, these are not taken into account quantitatively.

In this chapter, we present a Bayesian random-effects meta-analysis of the relative clause data (Sutton et al., 2012; Gelman et al., 2014). We began by testing a model written in the probabilistic programming language JAGS (Plummer, 2012) for the meta-analysis by fitting simulated data; this evaluation confirmed that the model can recover the true parameters. We then modelled the data. To anticipate the main result in this chapter, the posterior distribution of the parameter suggests that the posterior probability of the effect being positive, i.e., the probability of a subject relative advantage, is 0.76.

3.2 A random effect meta-analysis of the relative clause data

One way to conduct a meta-analysis is to conduct a so-called fixed-effects meta-analysis (Chen and Peace, 2013). This assumes that all the studies have a true effect θ. Thus, if the observed effects from i studies are $\hat{\theta}_i$, then the fixed-effects model is $\theta_i \sim Normal(\theta, \sigma^2)$.

If, however, it is more reasonable to assume that each study has a different θ, then one can conduct a so-called random-effects meta-analysis. This would assume that each study i has an underlying true mean θ_i that is generated from a normal distribution $Normal(\theta, \tau^2)$, and that each observed effect y_i is generated from $Normal(\theta_i, {s_i}^2)$, where s_i is the estimated standard error from study i. Thus, the random-effects meta-analysis has a new parameter, τ^2, that characterizes between-study variance. The fixed-effects meta-analysis is in fact just a special case of the random-effects model, under the assumption that $\tau = 0$.

In our case, a random-effects meta-analysis makes more sense because it is likely that there is significant heterogeneity the studies, since they were run under different conditions. Therefore, we first carried out a random-effects meta-analysis of the available data.[1] We turn to the description of this model next.

3.2.1 Model specification

The model was the following. Let y_i be the effect size in milliseconds in the i-th study, where i ranges from 1 to n (in this dissertation, $n = 15$). A positive sign of a value y_i indicates a subject relative advantage and a negative sign an object relative advantage. Let θ be the true (unknown) effect, to be estimated by the model. Let ${\sigma_i}^2$ be the true variance of the sampling distribution; each σ_i is estimated from the sample standard error from study i. The variance parameter τ^2 represents between-study variance.

Then, our model for n studies is:

$$\begin{aligned} y_i \mid \theta_i, {\sigma_i}^2 &\sim N(\theta_i, {\sigma_i}^2) \quad i = 1, \ldots, n \\ \theta_i \mid \theta, \tau^2 &\sim N(\theta, \tau^2), \\ \theta &\sim N(0, 100^2), \\ 1/\tau^2 &\sim Gamma(0.001, 0.001) \end{aligned} \tag{3.1}$$

Although we show a Gamma prior above for the between-study precision, there are several alternative plausible priors we can use for τ^2. We will

[1] An initial version of the random effects meta-analysis reported here appeared in Vasishth et al. (2013).

consider three priors (see Gelman (2006) for discussion on the choice of priors for variance components):

1. A Gamma prior on $1/\tau^2$: $1/\tau^2 \sim Gamma(0.001, 0.001)$.

2. A uniform prior on τ: $\tau \sim Uniform(0, 200)$.

3. A truncated normal prior on τ: $\tau \sim Normal(0, s^2)I(0,)$ for different standard deviations s. We choose $s = 200$ as the truncated $Normal(0, 200^2)$ covers plausible values of between-trial variance.

Although we do not expect the absolute effect to be larger than 30 ms for this particular research question regarding Chinese relatives, in psycholinguistics studies on relative clauses, plausible values of effect sizes can be assumed to range between -195 and 195 ms. This range is based on experience: effect sizes in psycholinguistics are rarely outside this range for relative clause studies. This is why we set a prior on θ to be $Normal(0, 100^2)$.

3.2.2 Simulation 1: Using simulated data to validate the JAGS code for random effects meta-analysis

First, we validated the random effects meta-analysis code by generating and fitting simulated data with known parameters. A function was written to generate data in the following manner; Table 3.1 shows example simulated data.

1. We chose, for $n = 15$ studies, the true effect $\theta = 15$, between-study variance $\tau^2 = 0.01^2$, standard error of each study fixed at $\sigma_i = 3$.

2. For each study $i = 1, \ldots, 15$, we sampled θ_i independently from $Normal(\theta, {\sigma_i}^2)$.

3. Then we generated observations $y_i \sim Normal(\theta_i, \sigma^2)$, where $\sigma^2 = 9$.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
θ_i	19	17	18	18	17	16	17	19	17	17	16	19	18	19	16
y_i	24	19	14	14	17	21	19	22	18	20	18	15	13	26	10

Table 3.1: An example of simulated data used in simulation 1. Shown are the means of the underlying generative normal distribution, and the effect in each study i.

We then derived $p(\theta \mid y_i)$, assuming the following likelihood and priors:

$$\begin{aligned} y_i \mid \theta_i, {\sigma_i}^2 &\sim N(\theta_i, {\sigma_i}^2) \quad i = 1, \ldots, n \\ \theta_i \mid \theta, \tau^2 &\sim N(\theta, \tau^2), \\ \theta &\sim N(0, 100^2), \\ \tau &\sim Uniform(0, 200) \end{aligned} \tag{3.2}$$

Then, we ran the JAGS model and sampled from the posterior distributions of θ, θ_i, τ. Four chains were run with a burn-in of 5000 iterations, and the total number of iterations was $20,000$ (a large number of iterations was run in order to ensure that the model converged). The Gelman-Rubin diagnostic (Gelman et al., 2014) (not shown) was used to confirm that we have successful convergence. This diagnostic essentially computes a statistic analogous to the F-statistic in ANOVA, by computing the ratio of the between-chain variance to within-chain variance. Thus, if the statistic has a value near 1, the chains are assumed to have mixed well, and the model is considered to have converged.

In Figure 3.1, we see the randomly generated effects of each study along with confidence intervals, the posterior distributions of each study with 95% credible intervals, and the posterior distribution of the effect given the randomly generated data. Figure 3.2 shows the marginal distributions of the parameters of interest in the analyses of simulated data.

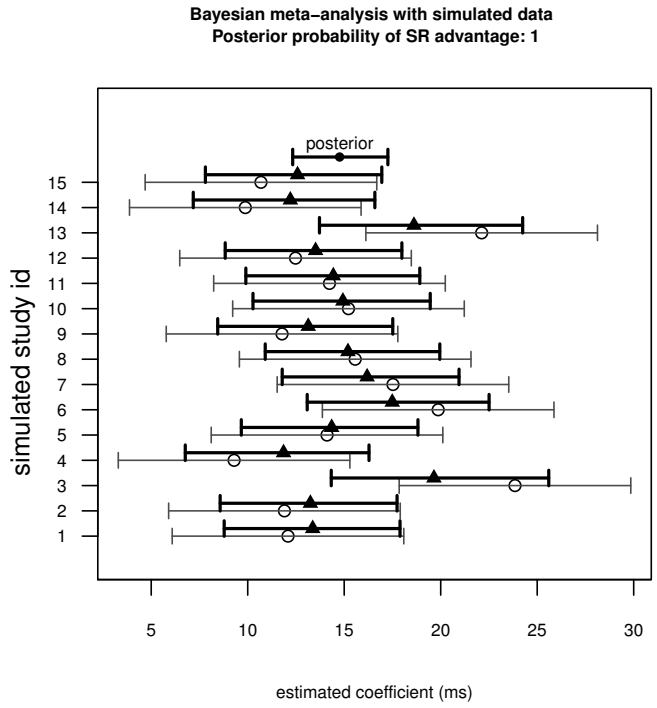


Figure 3.1: Simulation 1: Results of the random effects meta-analysis on simulated data. Shown are the means (circles) and 95% confidence intervals for each (randomly generated) study, the corresponding posterior means (triangles) and 95% credible intervals, and the posterior distribution mean and credible interval.

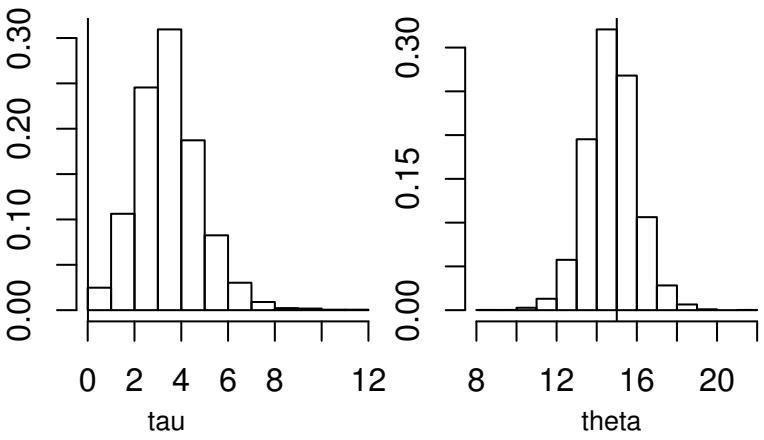


Figure 3.2: Simulation 1: Marginal posterior distributions of the main parameters of interest in the random effects meta-analysis of simulated data. The label tau refers to the between study variance; and theta is the posterior distribution of the true effect given the data. Also shown as vertical lines are the true values of $\theta = 15$ and $\tau = 0.01$.

3.2.3 Simulation 2: Sensitivity analysis of the random-effects meta-analysis

A second test of the validity of the code is to repeatedly fit randomly generated data using the above model, with different priors for the between-trial standard deviation τ. We repeatedly generated random data 20 times, and fitted the random-effects meta-analysis to each data-set. Figure 3.3 shows 95% credible intervals and medians of the two parameters (θ and τ), for each of the three priors for τ: Gamma(0.001,0.001) on $1/\tau^2$, Uniform(0,200) on τ, and a truncated normal with mean 0 and standard deviation 200.

Discussion of simulations 1 and 2

The simulation shows that the random-effects meta-analysis model can indeed recover the true θ; under repeated runs of the model, the true θ is contained within the 95% credible interval of the posterior distribution. When the prior for $1/\tau^2$ is Gamma(0.001,0.001), the posterior estimate for τ does fall with the range of plausible values implied by the posterior distribution. With the other priors, the posterior distribution is somewhat overestimated.

In conclusion, the JAGS code for the random effects meta-analysis seems to be performing as expected, giving us confidence that we can use it to study the observed data.

3.3 Random effects meta-analysis of the available data

Next, we describe the random effects meta-analysis of the 15 studies discussed in chapter 2. As above, four chains were run, with a burn-in period of 5000 iterations, and a total of 20,000 iterations. Model convergence was

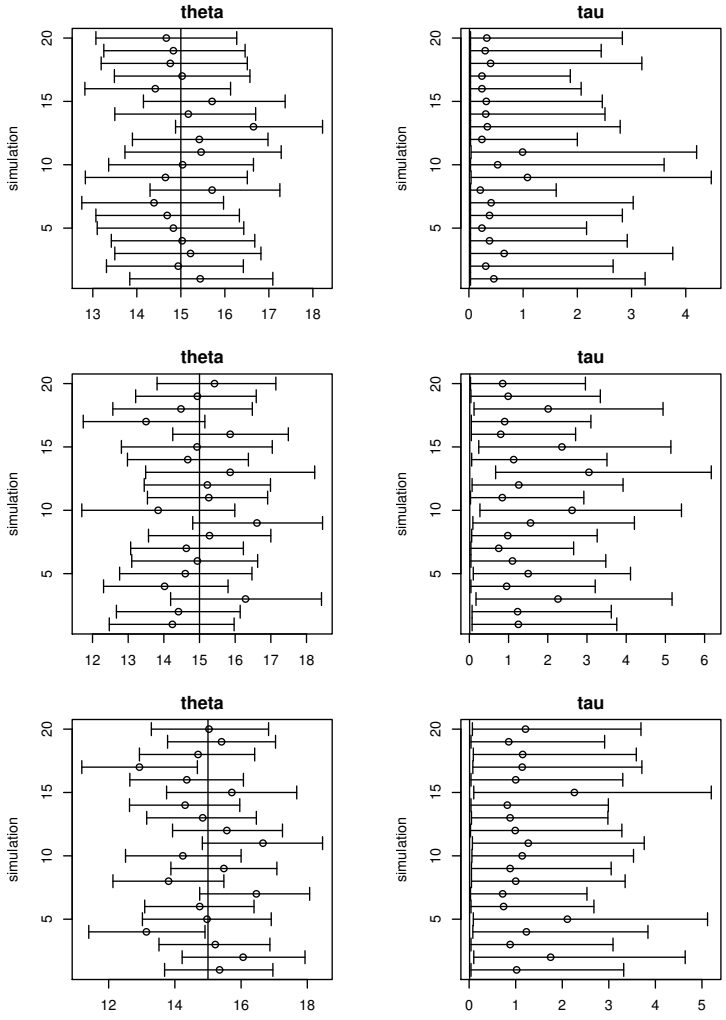


Figure 3.3: Simulation 2: Result of repeated JAGS model fits on randomly generated data, with a Gamma(0.001,0.001) prior used for between-trial precision (top two figures); a Uniform(0,200) prior for the between-trial standard deviation (middle two figures); and a truncated Normal prior for the between-trial standard deviation (bottom two figures). The true values of the parameters are shown as vertical lines.

checked visually, by plotting the trajectories of the chains, and by using the Gelman-Rubin diagnostic (Gelman et al., 2014). Convergence was successful for each parameter. Figure 3.4 summarizes the results of the meta-analysis, and Figure 3.5 shows the marginal posterior distributions of τ and θ. The posterior probability of a subject relative advantage is 0.78, with mean 16 and 95% credible intervals -29 and 59.

3.3.1 Discussion

The random-effects meta-analysis of the 15 studies shows that there is weak evidence in favour of the SR advantage; the posterior probability of an SR advantage is approximately 0.78. The posterior distributions of each of the individual studies is shifted closer to the grand mean; this is an instance of the shrinkage that is characteristic of hierarchical models (Gelman and Hill, 2007).

The meta-analyis thus provides some clarity about the state of the current evidence regarding this issue. However, as mentioned earlier, it is likely that many (if not all) of these studies have significant biases that could have skewed the results; this makes the posterior distribution difficult to interpret. Here, bias modelling is a useful alternative; if we can quantify different sources of bias across studies, then we can incorporate these sources of bias into the meta-analysis. We discuss bias modelling in the next chapter.

3.4 Conclusion

We carried out a random-effects meta-analysis of the Chinese relative clause data. The posterior distribution of the effect of interest shows that the probability of the parameter being positive is approximately 0.78, with

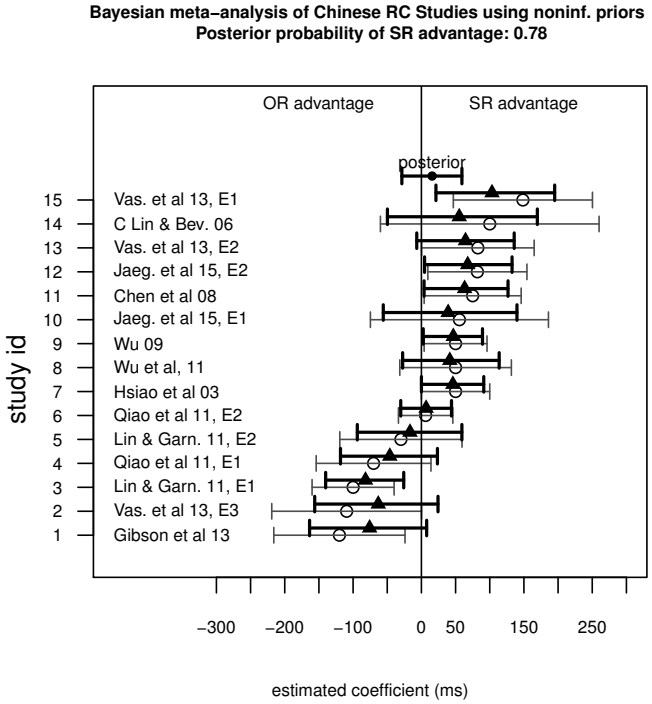


Figure 3.4: Results of random effects meta-analysis. Shown are the means (circles) and 95% confidence intervals for each study, the corresponding posterior means (triangles) and 95% credible intervals, and the posterior distribution mean and credible intervals.

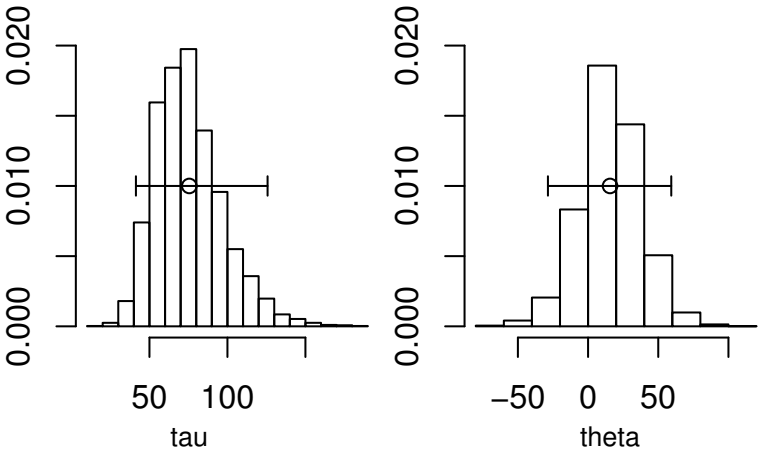


Figure 3.5: Posterior distributions of the main parameters of interest in the random effects meta-analysis; the error bars show the means and the bounds of the 95% credible intervals. The label tau refers to the between study variance; and theta is the posterior distribution of the true effect given the data.

mean 16 and 95% credible intervals -29 and 59. In other words, there seems to be a tendency towards a subject-relative advantage. We turn next to our attempt at modelling biases.

CHAPTER 4

BIAS MODELLING IN META-ANALYSIS

4.1 Introduction

The random-effects meta-analysis discussed in chapter 3 assumes that the estimated mean of each study is generated from some true underlying distribution with some unknown mean θ and some between-study variance τ^2. However, different sources of bias could be present in the studies. The term bias here refers to *systematic* (as opposed to random) error or deviation from the true value, which either leads to an overestimate or an underestimate.

One response to the issue of bias has been to assess the methodological quality of each study, but these assessments are often not taken into account quantitatively in the meta-analysis. Removing studies that are considered to have bias (Sterne et al., 2001) is also not desirable, as potentially useful information is lost. The conventional method used to adjust for biases is through sensitivity analyses (e.g., carrying out a meta-analysis will all the data, and then removing potentially biased studies to investigate whether conclusions change) and/or exploratory sub-group analyses (Higgins and Green, 2008; Moja et al., 2005). Greenland and O'Rourke (2001) have attempted to incorporate biases into the analysis include weighting the analyses by quality scores, but this has the disadvantage that it only modified the variance of the estimate, and assumes that the magnitude of the effect is correctly estimated.

In response to this issue, Eddy et al. (1990) proposed a Bayesian approach that explicitly models internal and external biases by incorporating subjective judgements about them. Spiegelhalter and Best (2003) also explicitly incorporated biases additively, by choosing distributions for parameters representing internal and external biases. The approach presented by Turner et al. (2008) (also see Thompson et al. (2011)) builds on these previous attempts. Turner and colleagues propose a simple and generalizable method for adjusting for biases in meta-analyses. In this chapter, we describe this method in detail.

4.2 Bias modelling: The approach taken by Turner et al. (2008)

Turner et al. (2008) define bias in a study along two dimensions, rigour and relevance, which are discussed next.

4.2.1 Two sources of bias: Lack of rigour and relevance

Rigour refers to the presence or absence of *internal bias*. It is a measure of how well the parameters of interest are estimated. Different types of internal bias have been identified in the literature, and Turner and colleagues list the following.

1. *Selection bias* occurs when there are systematic differences between comparison groups; in the Chinese RC problem, most of the issues have to do with the selection of experimental items rather than subjects. For example, the experiment design may introduce a confound that inflates reading times in one condition but not another.

2. *Performance bias* arises due to factors such as inadequate blinding. A common example of this in psycholinguistic studies is when distractor items, called fillers, are used to mask the experimental manipulation so that subjects don't develop a strategy for reading the target sentences. If these fillers do not have a range of syntactic structures, the subject may be able to easily detect the effect, leading to a reading strategy that does not reflect realistic processing.

3. *Attrition bias* is driven by systematic (i.e., non-random) differences between comparison groups when excluding data. In psycholinguistics, attrition bias occurs frequently due to selective removal of data. In the studies considered here, the Gibson and Wu (2013) data has an instructive example of such attrition bias: in the published analysis, one item was removed from the data-set post-hoc, and removing this single item leads to a statistically significant effect.

4. *Detection bias* refers to differences induced by inadequate blinding from those assessing the outcome. If the data analyst has a stake in the outcome of the experiment, then he/she could be highly motivated to somehow obtain a statistically significant effect. This is unfortunately the normal situation in psycholinguistics: the experimenter is usually aiming to provide evidence in favour of a theoretical point that they already believe in before they run an experiment. Thus, there is little hope for objectivity in the analysis.

5. *Other bias suspected* is an open-ended category, intended to cover any other biases not included in the above list.

Relevance, which refers to (the absence of) external bias, is defined with respect to the specific research question. Turner et al identify the following factors.

1. *Population bias* occurs when there are differences in age, sex, or health status of the idealized study participants; in the Chinese RC context, the main source of population bias is likely to come from experimental items rather than subjects, although some studies (such as Hsiao and Gibson (2003)) have unusually old populations, which can bias the effect. Another example of population bias in the Chinese relative clause case would be the situation where we want to know about processing differences in adults, but we have data on children.

2. *Intervention and control bias* refer to differences in delivery of the experimental manipulation (respectively, control). In the present study, this never occurs in any of the papers because all are planned experiments with within-subjects repeated measures designs. Consequently, we exclude these factors, noting only that for between-subject studies, these biases should be considered.

3. *Outcome bias* refers to differences in method of measurement of the idealized study compared to target. An example of outcome bias is the study by Qiao et al. (2012), which uses the highly non-standard maze task instead of an SPR or eyetracking study to investigate comprehension difficulty.

The extent of internal and external bias in a study is unknown, but can be elicited from experts. The effect these will generally have on the observed

estimates from the data will be to shift the mean and widen the confidence (or credible) interval.

4.3 Steps in bias identification

Turner and colleagues suggest the following steps in identifying bias.

1. Define the target question and the target experimental manipulation, including the population being studied, and the outcome of interest.

2. Define an *idealized version* of each source study and write down a mini-protocol that lists each component of the idealized study. The idealized study is defined as a repeat of the original study, but one having a design that has no sources of internal bias. In the idealized study, we define the population to be studied, the planned comparison, and the outcome that is planned to be measured. This information can be extracted from the Methods section of each paper under consideration.

3. Compare the details of the completed source study against the mini-protocol defined in the previous step.

4. Identify internal bias by comparing each idealized study with the target study.

5. Identify external bias by comparing each idealized study with the target study.

The implemention of Steps 4 and 5 is discussed next.

4.4 Steps for identifying internal and external bias

Table 4.1 shows a checklist that was used to identify sources of bias in each study. In this table, all biases are additive (see next section for an explanation). Turner et al had also considered whether each bias could be treated as proportional; a bias is proportional if it depends on the magnitude of the effect. Proportional biases are relevant in medical studies where patient participation or patient drop-outs might depend on how seriously ill the patient is. In psycholinguistics, subjects are generally unaware of the experimental manipulation, so proportional biases can be assumed to play no role here.

The expert assessors are asked to complete a checklist for each study. See appendix A on page 75 for examples of completed checklists.

4.5 Adjusting means and variances by incorporating biases

If there were no *internal* biases, the generating distribution would be

$$y_i \sim Normal(\theta_i, {s_i}^2) \tag{4.1}$$

where i indexes the study, θ_i is the true study-level effect such that $\theta_i \sim Normal(\theta, \tau^2)$, and ${s_i}^2$ is the variance for the sampling distribution of the mean of the i-th study.

If there were no *external* biases, $\theta_i = \theta$, where θ is the true underlying effect, which is assumed to have a true, unknown point value.

Next, we discuss the adjustments to each study's mean θ_i and variance ${s_i}^2$ as a function of the internal and external bias. As mentioned above, the adjustments

Bias
Internal Bias
Selection bias
Subjects in all conditions recruited from same populations?
Subjects recruited over same time periods?
Were inclusion and exclusion criteria clear?
Was randomization used?
Did the comparison conditions constitute a fair comparison (were they minimal pairs)?
Performance bias
Were subjects blinded?
Was the experimenter blinded?
Adequate concealment of experimental manipulation (adequate use of filler sentences to mask the experimental manipulation)?
Was the experimental method appropriate?
Attrition bias
Were any subjects excluded post-hoc?
Are the results likely to be affected by post hoc exclusions?
Detection bias
Was data analyst blinded?
Reading time measured accurately (appropriate software used, lab conditions)?
Was the statistical analysis appropriate?
Other bias suspected
Do you suspect other bias?
External Bias
Population bias
Were study subjects in idealized study drawn from population identical to target population, with respect to native speaker status?
Outcome bias
Was study outcome for idealized study identical to target outcome?

Table 4.1: Checklist for identifying internal and external bias.

assume that biases are independent of the magnitude of the effect; these are called additive biases by Turner and colleagues.

4.5.1 Model specification

The superscripts I and E represent internal and external biases respectively.

Internal bias

If there are j multiple independent sources of internal bias, where $j = 1, \ldots, J^I$, we will write that ${\delta^I}_{ij}$ is the effect of bias source j on the estimated effect in study i.

${\delta^I}_{ij}$ is unknown, and our uncertainty about ${\delta^I}_{ij}$ is expressed by the distribution ${\delta^I}_{ij} \sim N({\mu_{ij}}^I, ({\sigma_{ij}}^I)^2)$ (or perhaps a t-distribution with low degrees of freedom; we do not consider this option in the dissertation).

So our goal is to elicit values for these parameters, ${\mu^I}_{ij}, ({\sigma_{ij}}^I)^2$.

The **total internal bias** in the ith study is

$$ {\delta^I}_i \sim N({\mu_i}^I, ({\sigma_i}^I)^2) \tag{4.2} $$

where ${\mu_i}^I = \sum_{j=1}^{J^I} {\mu_{ij}}^I$, and $({\sigma_i}^I)^2 = \sum_{j=1}^{J^I} ({\sigma_{ij}}^I)^2$.

These biases are assumed to influence the θ additively:

$$ y_i \sim N(\theta_i + {\mu_i}^I, {s^2}_i + ({\sigma_i}^I)^2) \tag{4.3} $$

The term θ_i is explained below (equation 4.6).

External bias

Assuming multiple independent sources $j = 1, \ldots, J^E$ of external bias, we can write ${\delta_{ij}}^E$ to represent the jth source of external bias in study i. Similar to the

internal bias case, each study is assumed to have external bias ${\delta_{ij}}^E$, defined by the distribution:

$${\delta_{ij}}^E \sim N({\mu_{ij}}^E, ({\sigma_{ij}}^E)^2) \tag{4.4}$$

Turner and colleagues included a variance parameter τ^2 to represent unexplained between-study heterogeneity. If, hypothetically, we could adjust internal and external biases perfectly, there would be no remaining heterogeneity, and τ^2 would have value 0.

After we have elicited ${\mu_{ij}}^E$ and $({\sigma_{ij}}^E)^2$ for each study, the total external bias in the i-th study would be

$${\delta_i}^E \sim N({\mu_i}^E, ({\sigma_i}^E)^2) \tag{4.5}$$

where ${\mu_i}^E = \sum_{j=1}^{J^E} {\mu_{ij}}^2$ and $({\sigma_i}^E)^2 = \sum_{j=1}^{J^E} ({\sigma_{ij}}^E)^2$.

The external bias model can then be written as:

$$\theta_i \sim N(\theta + {\mu_i}^E, \tau^2 + ({\sigma_i}^E)^2) \tag{4.6}$$

Thus, assuming no internal biases, we have

$$y_i \sim N(\theta + {\mu_i}^E, {s_i}^2 + \tau^2 + ({\sigma_i}^E)^2) \tag{4.7}$$

Taking both internal and external bias into account

If we include both sources of bias, the observed effect in each study is:

$$y_i \sim N(\theta + {\mu_i}^I + {\mu_i}^E, {s_i}^2 + ({\sigma_i}^I)^2 + \tau^2 + ({\sigma_i}^E)^2) \tag{4.8}$$

This is a random-effects meta-analysis with the mean adjusted for biases. We will refer to this model as a bias-adjusted random-effects meta-analysis.

4.5.2 Bias elicitation procedure: The Sheffield Elicitation Framework v2.0

The procedure adopted by Turner and colleagues is as follows (they assume multiple assessors):

1. Each assessor completes checklists (see Table 4.1) for sources of bias. This requires reading the Methods and Results section of each paper.

2. The assessors then meet to discuss their checklists and the papers, and arrive at a consensus regarding the sources of bias.

3. Each assessor writes down a 95% confidence interval for each source of bias, using elicitation scales. This is done by each assessor independently.

4. Distributions from the assessors are pooled. Turner and colleagues carry out the pooling for each bias by taking the median of the assessors' means, and the medians of the standard deviations.

5. If distributions for a particular paper and bias are extremely divergent among the assessors, the assessors consult with each other to arrive at a consensus.

In the present work, instead of the above procedure, we adopted the Sheffield Elicitation Framework v2.0 (Oakley and O'Hagan, 2010), available from the website http://www.tonyohagan.co.uk/shelf/.

This framework, referred to hereafter as SHELF, has the advantage of providing a detailed set of instructions and a fixed procedure for eliciting distributions. It also

provides detailed guidance on documenting the elicitation process, thereby allowing a full record of the elicitation process to be created. The SHELF procedure works as follows. There is a facilitator and an expert (or a group of experts; we will consider the single expert case here).

1. A pre-elicitation form is filled out by the facilitator in consultation with the expert. This form sets the stage for the elicitation exercise and records some background information, such as the nature of the expertise of the assessor.

2. Then, an elicitation method is chosen. We chose the quartile method, so we discuss this as an example. The expert first decides on a lower and upper limit of possible values for the quantity to be estimated; this minimizes the effects of the "anchoring and adjustment heuristic" (O'Hagan et al., 2006), whereby experts tend to anchor their subsequent estimates of quartiles based on their first judgement of the median. Following this, a median value is decided on, and lower and upper quartiles are elicited. The SHELF software displays these quartiles graphically, allowing the expert to adjust them at this stage if necessary. It is important for the expert to confirm that, in his/her judgement, the four partitioned regions that result have equal probability.

3. The elicited distribution is then displayed as a density (several choices of probability density functions are available, but we will use the normal); this serves to give feedback to the expert. The parameters of the distribution are also displayed. Once the expert agrees to the final density, the parameters can be considered the expert's judgement regarding the prior distribution of the bias.

4.5.3 The expert assessor

In the present dissertation, we have only one assessor available (Shravan Vasishth), who is also the author of the dissertation. It may be possible to find a second assessor at a later stage (after completion of the dissertation). The main limitation is that the assessor needs to be an expert in the area of Chinese relative clause processing, and there are very few people available with this knowledge. Also, it would be preferable to have one unbiased assessor, and a third assessor who has reason to believe in the object relative advantage.

The assessor in this dissertation has been working on models of language comprehension for 15 years at the time of writing, and is an expert on relative clause processing. The assessor has published two articles on Chinese relative clauses Vasishth et al. (2013); Jäger et al. (2015), and has conducted research on relative clauses in other languages, such as English (Bartek et al., 2011), German (Vasishth et al., 2011; Frank et al., 2015), Hindi (Vasishth, 2003; Vasishth and Lewis, 2006; Husain et al., 2014), and Persian (Safavi et al., 2015). He could be considered to have a bias in favour of the subject-relative advantage (positive sign for the parameter of interest), as his published work on Chinese supports that position.

4.6 Sensitivity analysis

In principle, the influence of variability in opinion among the assessors can be assessed by carrying out several analyses: (i) a bias-adjusted random effects meta-analysis using each assessor's elicited values separately, and (ii) another one pooling the assessors' values. In our case, since there is only one assessor, we will check the effect of including bias by comparing posterior distributions with and without bias modelling.

4.7 Conclusion

In this chapter, we discussed internal and external biases, and explained the framework for identifying and quantifying biases as developed by Turner and colleagues. We also presented the details of the SHELF framework that we use in chapter 5 to elicit biases.

CHAPTER 5

BIAS MODELLING OF THE CHINESE RELATIVE CLAUSE DATA

5.1 Introduction

In this chapter, we use the SHELF framework to elicit values for the biases in five of the fifteen studies, and then use these values in a Bayesian random-effects meta-analysis. We restrict the bias modelling to five studies due to time constraints. Before we fit the model (written in JAGS), we test the code using simulated data; this serves to establish the fact that the model is able to recover the key parameters of interest. Then, we fit the data using the JAGS code; as a baseline, we also compute the estimates for the effect using the analytical formulas presented by Turner and colleagues. The bias modelling shows that, after taking biases into account in five of the fifteen studies, the evidence for a subject relative advantage is stronger compared to the standard meta-analysis.

5.2 Some definitions

It will be useful to define some terms that we will use in this chapter. We will refer to the **effect** as the difference in reading itme (in milliseconds) between subject and object relatives in Chinese, measured at the head noun. A positive sign on the effect signals a **subject relative advantage** (or SR advantage), and a negative sign signals an **object relative advantage** (or OR advantage).

The **target question** is: Are subject relatives easier or more difficult to process than object relatives, as measured by differences in reading time at the head noun? The **target experimental manipulation** is a standard repeated measures Latin square design that compares reading times in milliseconds at the head noun in subject versus object relative clauses. The **population being studied** is unimpaired adult native speakers of Mandarin Chinese. The **outcome of interest** is the difference in reading time between object and subject relatives at the head noun.

We begin by defining the internal and external biases for the five studied considered in this dissertation, how these biases could have been overcome in an idealized study, and the possible effect they could have on the effect.

5.3 Target studies vs idealized studies

For each study, we identified sources of internal and external bias, and elicited a prior distribution for each bias using the SHELF framework. Below, we present the details of this bias-identication process for one study; the details for the other studies are described in the appendix.

5.3.1 Example: Vasishth et al 2013 Expt 3

Internal biases

1. **Selection**:

    (a) Subjects in all conditions recruited from same populations? *This is a within-subjects design, as is standard in psycholinguistics. Thus, by definition, subjects in all conditions come from the same population. No improvement is necessary in an idealized design.*

    Effect on study: None expected.

(b) Subjects recruited over same time periods? *The experimenter did not make this clear, but in the absence of any other information, we can assume that the experiment was not done over an extended period.*

Effect on study: None expected.

(c) Were inclusion and exclusion criteria clear? *All participants are stated to be native speakers of Mandarin. No improvement is necessary.*

Effect on study: None expected.

(d) Was randomization used? *It is not clear how the list was chosen for each incoming subject. Ideally, each incoming subject should have been assigned to a separate list; not doing this could lead to an over- or underestimate of the effect.*

Effect on study: None expected.

(e) Did the comparison conditions constitute a fair comparison (were they minimal pairs)? *This experiment has the following possible confound. Charles Lin (Effect of thematic order on the comprehension of Chinese relative clauses. Lingua, 140, 180-206, 2014) has pointed out that the context sentences in this experiment could have made subject relatives harder to process, since the thematic roles are reversed between the context and target sentence in subject (but not object) relatives. This potential confound is also present in the original Gibson and Wu 2013 study that this experiment attempted to replicate. An idealized design would have thematic roles appearing in the same order as in the target sentence. The confound is likely to bias the effect to be an overestimate; in fact, the effect could entirely be due to the confound.*

Effect on study: The observed effect (which had a negative sign) could

be entirely due to the thematic role reversal; i.e., the true effect could be zero or even have a positive sign.

2. **Performance**:

   (a) Were subjects blinded? *Yes.*

   Effect on study: None expected.

   (b) Was the experimenter blinded? *No. The experimenter was a co-author of the paper, Qiang Li, and he was aware of all aspects of the experiment.*

   Effect on study: None expected.

   (c) Adequate concealment of experimental manipulation (adequate use of filler sentences to mask the experimental manipulation)?

   Not clear. An idealized version would have a range of syntactic constructions as fillers, to prevent the subject from detecting that the experiment is about relative clauses.

   Effect on study: None expected.

   (d) Was the experimental method appropriate? *Yes. This was a standard self-paced reading design.*

   Effect on study: None expected.

3. **Attrition**:

   (a) Were any subjects excluded post-hoc? *No.*

   Effect on study: None expected.

   (b) Are the results likely to be affected by post hoc exclusions? *No.*

   Effect on study: None expected.

4. **Detection**:

   (a) Was data analyst blinded? *No; the analyst was Shravan Vasishth.*

   Effect on study: Lack of blinding can lead to biases in the analyses (p-value hacking, or garden-of-forking-paths effects, Gelman and Loken (2013)). However, since we ran a pre-determined analysis, no bias is expected.

   Effect on study: None expected.

   (b) Reading time measured accurately (appropriate software used, lab conditions)? *Yes. Standard self-paced reading software was used in a laboratory setting (Linger: http://tedlab.mit.edu/∼dr/Linger/).*

   Effect on study: None expected.

   (c) Was the statistical analysis appropriate? *Yes. A linear mixed model was fitted for estimatation and inference.*

   Effect on study: None expected.

5. **Other**:

   (a) Do you suspect other bias? *No.*

   Effect on study: None.

External biases

In each of the studies, the **target population** is normal, cognitively unimpaired adult native speakers of Chinese (any variety, but all studies considered happen to be about Mandarin Chinese), but psycholinguistic experiments typically have university students (usually undergraduates) who are relatively young; the **comparison** of interest is subject vs object relative clause processing times

at the head noun; the **outcome of interest** is the difference in reading times at the head noun.

Population All participants are stated to be native speakers of Mandarin at Dalian University of Technology, and were undergraduate students at this university. This matches the target population.

Effect on study: None.

Outcome The measurement was done using self-paced reading. This is an appropriate method for the research question.

Effect on study: None.

5.3.2 Example (continued): Bias elicitation procedure for Vasishth et al 2013, Expt 3

Here, we provided a detailed example of how the biases were elicited for the above study. For the other four experiments, see the SHELF framework worksheets in appendix C on page 100.

1. First, the facilitator and the expert (Shravan Vasishth) filled out the pre-elicitation form. Two major issues became clear from the pre-elicitation form. The first is that the expert is a stakeholder in the results of the meta-analysis: he has published papers on Chinese relative clauses that have predominantly shown a subject-relative advantage. However, his own theoretical work (Lewis and Vasishth, 2005) predicts an object relative advantage; so, on balance, one could in principle consider him unbiased. The second, more critical issue is that the expert has no quantitative feel for effect sizes in his area, since most of the work done in psycholinguistics involves significance testing; the only issue of interest usually is whether an effect is

statistically significant or not. If multiple experiments are done and all show a negative sign on the coefficient, but all are statistically non-significant, one would usually conclude that the true value of the parameter is 0.

2. Then, the elicitation was carried out using the quartile method. In this particular paper (Vasishth et al, 2013), the only source of bias is selection bias, as discussed above. Therefore, we discuss the elicitation process only for this case. The expert first identified upper and lower bounds on the possible values of each bias as -200 and 0; the justification for these bounds was that the bias is expected to cause negative reading times (longer reading times in subject relatives vs object relatives), and the absolute values of effects are only rarely higher than 200 ms. Then, a median value of -60 ms, with upper and lower quartiles were -55 and -140 were elicited. After this, the expert was shown the four equi-probable regions, and once he had judged them as having equal probability, the density plot was shown. This revealed that there was too much probability in the positive region. This led to a revision of the median to -90 and the 0.05 and 0.95th intervals were revised to -73 and -47. The final distribution settled on was $Normal(-90, 25^2)$. All other biases are assumed to come from a normal distribution with mean 0 and variance 0.01^2 (i.e., effectively no bias).

5.4 Bias modelling

The model defined in section 4.5.1 was implemented in JAGS. To summarize the model, if we include both sources of bias, the observed effect y_i, where $i = 1, \ldots, n$, n the number of studies, is modelled as:

$$y_i \sim N(\theta + {\mu_i}^I + {\mu_i}^E, {s_i}^2 + ({\sigma_i}^I)^2 + \tau^2 + ({\sigma_i}^E)^2) \tag{5.1}$$

The goal is to derive the posterior distribution of θ.

5.5 Simulation 3: The standard random-effects meta-analysis cannot recover parameters in biased data

Before we carried out the bias modelling, we first established that a standard random-effects meta-analysis would be unable to recover the true effects. In order to do this, we generated simulated data again, but with internal and external biases included. The goal here was to check whether bias shifts the posterior distribution away from the true value of $\theta = 15$. In other words, due to the biases, the posterior should no longer include the true value. In addition, the between-study standard deviation τ should also become larger. If these two changes in the posterior distributions of θ and τ are seen, then we have established that if there is bias in the data, the ordinary random effects meta-analysis cannot accurately recover the true mean.

A function was written to generate data as follows. Example simulated data are shown in Table 5.1.

1. We chose, for $n = 15$ studies, the true effect $\theta = 15$, between-study variance $\tau^2 = 0.01^2$, standard error of each study fixed at $\sigma_i = 3$.

2. For each study $i = 1, \ldots, 15$, we set priors for external bias E_{ij}, j=1,2 to be $Normal(1, 0.5)$, and then sampled θ_i independently from $Normal(\theta + \sum_j E_{ij}, \tau^2 + \sum_j {ESD_{ij}}^2)$, where ${ESD_{ij}}^2 = 0.5$ is the variance of bias j.

3. Then, for each study i, we sampled internal bias means I_{ik}, k=1,...,5 from $Normal(1, 0.5)$.

4. Then we generated observations $y_i \sim Normal(\theta_i + \sum_k I_{ik}, \sigma^2 + \sum_j {ISD_{ik}}^2)$, where $\sigma^2 = 9$ and ${ISD_{ik}}^2$ is the variance associated with internal bias.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
θ_i	18	15	17	17	18	18	17	19	17	16	16	16	17	18	17
y_i	14	27	28	22	17	24	22	21	17	18	10	18	21	14	24

Table 5.1: Example biased data used in simulation 3.

We fit the standard random effects meta-analysis and plot the posterior distributions. In Figure 5.1, we see the randomly generated effects of each study along with confidence intervals, the posterior distributions of each study with 95% credible intervals, and the posterior distribution of the effect given the randomly generated (biased) data. Figure 5.2 shows the marginal distributions of the parameters of interest.

5.5.1 Discussion of simulation 3

The simulation shows that in the presence of internal and external bias, the standard random effects meta analysis is unable to recover the posterior distributions. Next, we attempt to model simulated biased data using bias-adjusted random-effects meta-analysis.

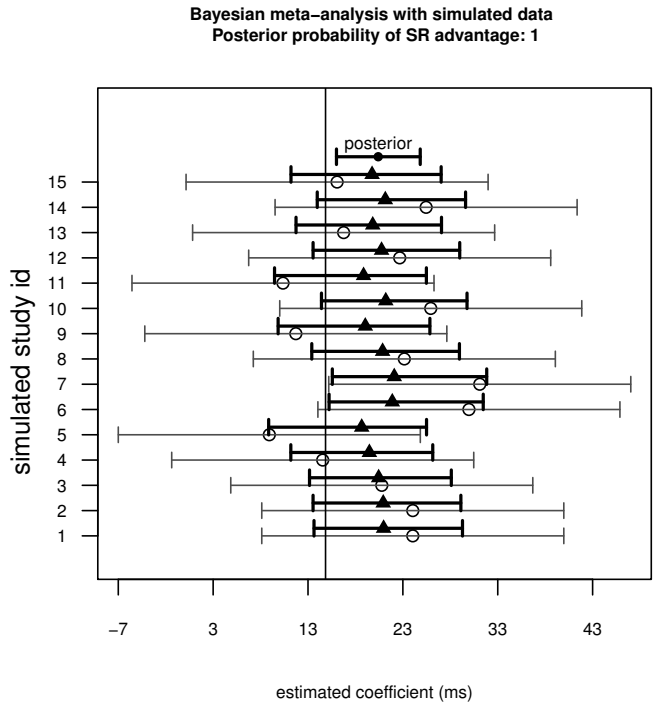


Figure 5.1: Simulation 3: Results of the random-effects meta-analysis model fit to biased data. The true value of $\theta = 15$ is marked by the vertical line. It is clear that the standard meta-analysis is over-estimating the true value.

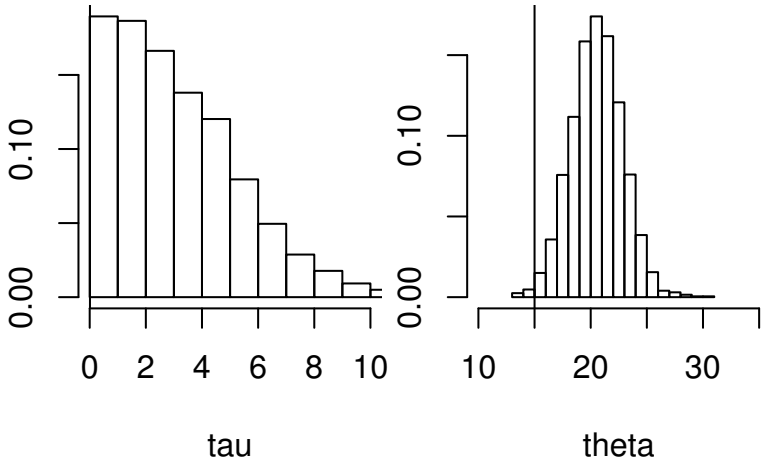


Figure 5.2: Simulation 3: Marginal posterior distributions of the main parameters of interest in a standard random effects meta-analysis using simulated data. The label tau refers to the between study variance; and theta is the posterior distribution of the true effect given the data. Also shown as vertical lines are the true values of θ and τ.

5.6 Simulation 4: Validating the JAGS code for bias modelling using simulated data

To validate the JAGS code that we developed for bias modelling, we first generated biased data for 15 studies with known means and variances for five internal sources of bias and two external sources of bias. The goal was to determine whether the JAGS model can recover the true effects given accurate values for the biases.

The same function as in simulation 3 was used to generate data. Example simulated data are shown in Table 5.2. We repeat a description of the procedure here for convenience.

1. We chose, for $n = 15$ studies, the true effect $\theta = 15$, between-study variance $\tau^2 = 0.01^2$, standard error of each study fixed at $\sigma_i = 3$.

2. For each study $i = 1, \ldots, 15$, we set priors for external bias E_{ij}, j=1,2 to be $Normal(1, 0.5)$, and then sampled θ_i independently from $Normal(\theta + \sum_j E_{ij}, \tau^2 + \sum_j {ESD_{ij}}^2)$, where ${ESD_{ij}}^2 = 0.5$ is the variance of bias j.

3. Then, for each study i, we sampled internal bias means I_{ik}, k=1,...,5 from $Normal(1, 0.5)$.

4. Then we generated observations $y_i \sim Normal(\theta_i + \sum_k I_{ik}, \sigma^2 + \sum_j {ISD_{ik}}^2)$, where $\sigma^2 = 9$ is the standard error of each study, and ${ISD_{ik}}^2$ is the variance associated with the bias.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
θ_i	18	15	17	17	18	18	17	19	17	16	16	16	17	18	17
y_i	14	27	28	22	17	24	22	21	17	18	10	18	21	14	24

Table 5.2: Example biased data used in simulation 4.

We then derived $p(\theta \mid y_i)$, assuming the following likelihood and priors:

$$\begin{aligned}
\theta_i \mid \theta, E_{ij}, \tau^2, {ESD_{ij}}^2 &\sim Normal(\theta + \textstyle\sum_j E_{ij}, \tau^2 + \textstyle\sum_j {ESD_{ij}}^2) \\
y_i \mid \theta_i, I_{ik}, {\sigma_i}^2 + {ISD_{ik}}^2 &\sim Normal(\theta_i + \textstyle\sum_k I_{ik}, {\sigma_i}^2 + \textstyle\sum_k {ISD_{ik}}^2) \\
E_{ij} \text{ and } I_{ik} &\sim N(m = 1, v = 0.5) \\
\tau &\sim Uniform(0, 200) \\
\theta &\sim Normal(0, 100)
\end{aligned} \tag{5.2}$$

As a sensitivity analysis, we used three priors for the between-trial standard deviation τ: a Gamma(0.001,0.001) prior for $1/\tau^2$, a Uniform(0,200) prior on the standard deviation τ, and a truncated normal $Normal(0, 200^2)$ on τ. Figure 5.3 shows medians and 95% credible intervals of the posterior distributions of the two parameters, θ and τ, in repeated random sampling of biased data (20 runs) using these three priors. Figure 5.4 shows the posterior distributions of the individual studies and of the effect θ, and Figure 5.5 shows the marginal distributions of the parameters of interest using uniform priors on τ. We see that the model is able to recover the θ parameter, but only the Gamma prior on the precision of τ does a reasonable job of recovering the true value of τ. Although not shown, when the simulation was rerun with a much higher value of $\tau = 15$ (instead of 0.01), only the uniform and truncated normal prior on τ led to a posterior distribution that was able to cover the true value of τ; the Gamma prior greatly underestimated the mean for τ.

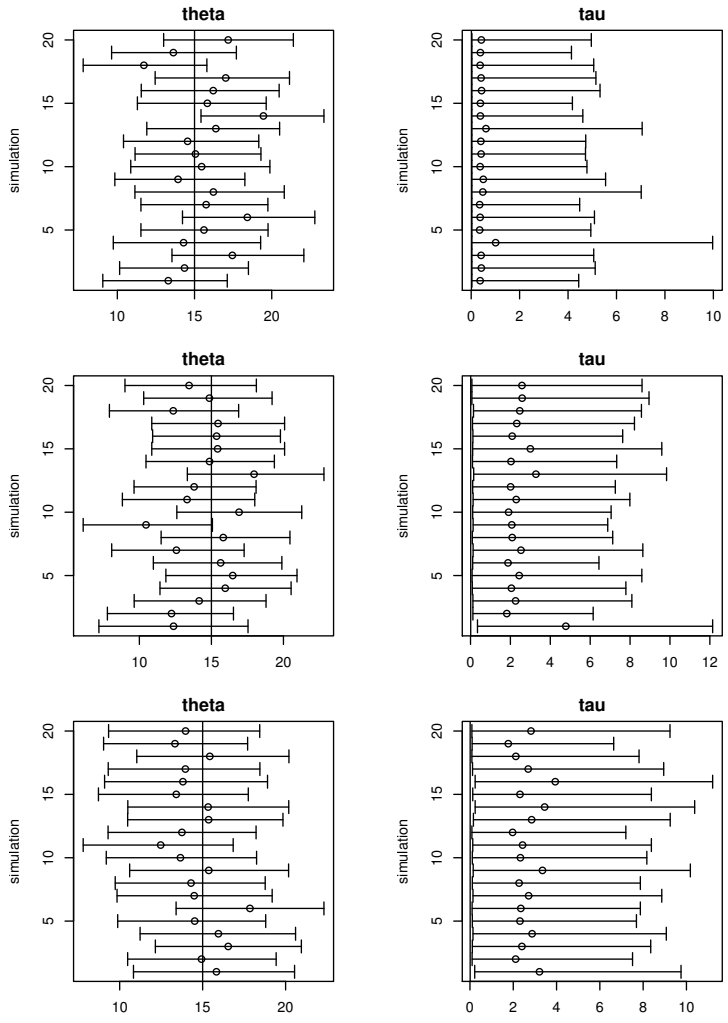


Figure 5.3: Simulation 4: Result of repeated JAGS model fits on bias modelling of randomly generated biased data, with a Gamma(0.001,0.001) prior used for between-trial precision (top panel); a Uniform(0,200) prior for the between-trial standard deviation (middle panel); and a truncated Normal prior for the between-trial standard deviation (bottom panel). The true values of the parameters are shown as vertical lines.

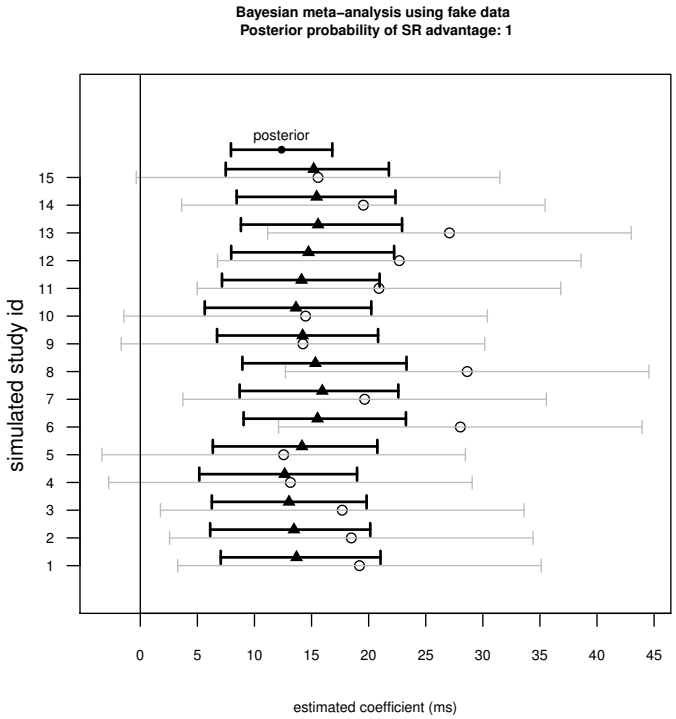


Figure 5.4: Simulation 4: Results of bias modelling using a random effects meta-analysis and simulated data. Shown are the means (circles) and 95% confidence intervals for each (randomly generated) study, the corresponding posterior means (triangles) with 95% credible intervals of the individual studies, and the posterior distribution mean with credible intervals of the effect.

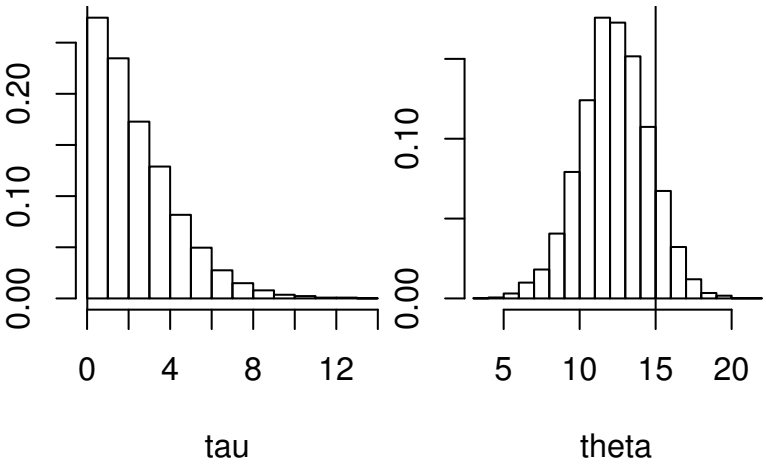


Figure 5.5: Simulation 4: Marginal posterior distributions of the main parameters of interest in the fake data random effects meta-analysis. The label tau refers to the between study variance; and theta is the posterior distribution of the true effect given the data. Also shown as vertical lines are the true values of θ and τ.

5.6.1 Discussion of simulation 4

The simulation shows that the model is able to recover the θ and τ parameters, in the sense that that the true values are covered by the posterior distributions of these parameters. This gives us confidence that we can use the model for our available data. The sensitivity analysis using different priors for τ suggests that a uniform or truncated normal prior might be a better choice than a Gamma(0.001,0.001) prior on precision $1/\tau^2$, as the Gamma prior consistently underestimates τ for larger values of this parameter. We will therefore use the uniform prior in our modelling: $\tau \sim Uniform(0, 200)$.

5.7 Analytical computation of an estimate of the true effect

Turner and colleagues also provide formulas for analytical calculation of the estimate of the true effect. These formulas are computed as follows:

1. The inverse variance estimator of θ is computed as follows. Given study i, define $\alpha_i = \frac{{s_i}^2}{{s_i}^2+{(\sigma^I)_i}^2}$, and $\gamma_i = \frac{\hat{\tau}^2}{\hat{\tau}^2+{(\sigma^E)_i}^2}$. Turner and colleagues explain these two quantities in the following manner. The term α_i can be interpreted as a quality weight for rigour, and represents the proportion of within-subject variability that is unrelated to internal biases (Spiegelhalter and Best, 2003). Highly rigorous studies would have α_i approximating 1, and less rigorous studies would have a lower value; a lower value of α_i has the consequence that it is downweighted in the analysis. Similarly, γ_i is a relevance weight, and represents the proportion of between-study variability that is unrelated to external biases (Spiegelhalter and Best, 2003). Studies that are highly relevant, i.e., without much bias, would have γ_i near 1, and studies that are less relevant, i.e., those having a lower γ_i, would be downweighted.

Then:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} \left[\frac{y_i - {\mu_i}^I - {\mu_i}^E}{{s_i}^2/\alpha_i + \hat{\tau}^2/\gamma_i} \right]}{\sum_{i=1}^{n} ({s_i}^2/\alpha_i + \hat{\tau}^2/\gamma_i)^{-1}} \tag{5.3}$$

with approximate standard error (based on the central limit theorem):

$$SE(\hat{\theta}) = \sqrt{\frac{1}{\sum_{i=1}^{n} ({s_i}^2/\alpha_i + \hat{\tau}^2/\gamma_i)^{-1}}} \tag{5.4}$$

2. An estimate of τ is computed using a method-of-moments estimate based on a heterogeneity statistic Q (DerSimonian and Laird, 1986) defined as

$$Q = \sum_{i=1}^{n} w_i (y_i - {\mu_i}^I - {\mu_i}^E - \hat{\theta}_F)^2 \tag{5.5}$$

where

$$\hat{\theta}_F = \frac{\sum_{i=1}^{n} w_i (y_i - {\mu_i}^I - {\mu_i}^E)}{\sum_{i=1}^{n} w_i} \tag{5.6}$$

and

$$w_i = ({s_i}^2 + (\sigma^I)_i{}^2 + (\sigma^E)_i{}^2)^{-1} \tag{5.7}$$

3. A moment estimate for τ^2 is

$$\hat{\tau}^2 = \frac{Q - (k-1)}{\sum_{i=1}^{n} w_i - \sum_{i=1}^{n} {w_i}^2 / \sum_{i=1}^{n} w_i} \tag{5.8}$$

If $\hat{\tau}^2$ is negative, then the estimate is set to 0; this is because negative values represent the case where within-study variance is larger than between-study variance, which Turner et al do not consider plausible.

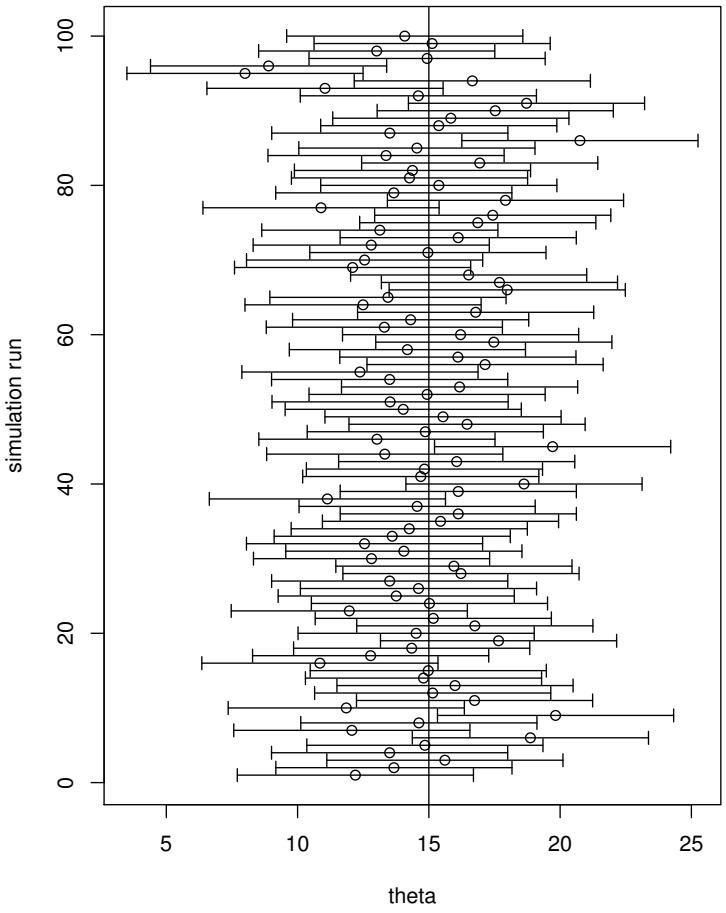


Figure 5.6: Results of simulation for evaluating the analytical formulas of Turner and colleagues. We see that the true value of $\theta = 15$ is contained in most of the intervals, confirming that the formula can estimate θ correctly.

In order to confirm that this analytical approach recovers the true effect of θ, we simulated data with internal and external biases repeatedly 100 times, and calculated the estimates of θ using the above formulas. The simulated data were generated as described earlier (page 59). The results of this simulation are shown in Figure 5.6. We see that the true value of θ is included in most of the 95% confidence intervals of the repeated samples. This validates the analytical approach and also provides us with a baseline to compare the JAGS model results with.

5.8 Bias modelling of the relative clause data

Study	Paper	Type	Bias	Mean	SD
1	GW13	Internal	Selection	-107	64
1	GW13	Internal	Attrition	-25.5	15.8
2	Vas13E3	Internal	Selection	-90	25
4	QiaoE1	Internal	Other	-50	31
4	QiaoE1	External	Outcome	-25	17
6	QiaoE2	Internal	Other	-51	31
6	QiaoE2	External	Outcome	-55.6	33.6
7	HG03	Internal	Other	37.4	26.5

Table 5.3: The elicited biases for the five studies. All categories of bias not shown were assessed by the expert to have no bias.

Model	bias-adjusted studies	lower	mean	upper	$Pr(\theta > 0)$
Standard	-	−28	16	61	0.77
Bias	1 (Gibson and Wu, 2013)	−17	24	65	0.88
Bias	2 (Vasishth et al., 2013)	−23	20	62	0.84
Bias	4 (Qiao et al., 2012) E1	−25	20	63	0.83
Bias	6 (Qiao et al., 2012) E2	−27	19	65	0.8
Bias	7 (Hsiao and Gibson, 2003)	−32	13	57	0.73
Bias	1,2,4,6,7	−4	33	72	0.96
Analytical	1,2,4,6,7	16	37	57	1

Table 5.4: The posterior probability of the effect of incorporating bias modelling for each study separately, compared to the standard random-effects meta-analysis. For each model, the mean and the bounds of the 95% credible interval are shown. Also shown (penultimate row) is the bias model with all five studies included (the remaining studies are assumed to have no bias). The final row shows the analytical calculation using Turner et al's formulas.

Having estimated the biases for the five studies (see section 5.3 and the appendix for details, and Table 5.3 for a summary of the elicited values), we fit

the bias-adjusted model. We assume here that the remaining 10 studies have no bias; eventually, we intend to determine biases for the remaining studies. Table 5.4 shows the effect that adjusting for bias in each of the five studies has on the posterior distribution given the 15 studies. This table also shows the posterior distribution of the standard meta-analysis as a baseline comparison, and the bias-adjusted model with all studies included. Finally, the table also shows the estimate of θ using the analytical formulas provided by Turner et al. In Figure 5.7, we see the posterior distribution of the effect θ and of each of the individual studies. The main outcome of the bias-adjusted model is that the posterior distribution of the parameter of interest is now more clearly positive (posterior probability of being positive: 0.96) than in the standard random-effects meta-analysis (posterior probability of being positive: 0.77). The analytical calculation using the formulas provided by Turner and colleagues gives much narrower bounds for the 95% confidence interval for $\hat{\theta}$.

5.9 Discussion

The analysis shows that taking the biases into account results in a positive value of the parameter, with a posterior probability of being greater than zero being 0.96. This suggests that, assuming that the bias adjustments have some validity, the evidence for a subject-relative advantage is stronger than was apparent when we did the standard random-effect meta-analysis. The theoretical implication is that we have evidence here against the dependency distance explanation of Chinese relative clause processing suggested by Hsiao and Gibson (2003).

It is worth stating here that the bias modelling presented here only serves as a tentative proof of concept; it would be misleading to draw inferences from the above attempt. A more convincing analysis would involve several experts' judgements on

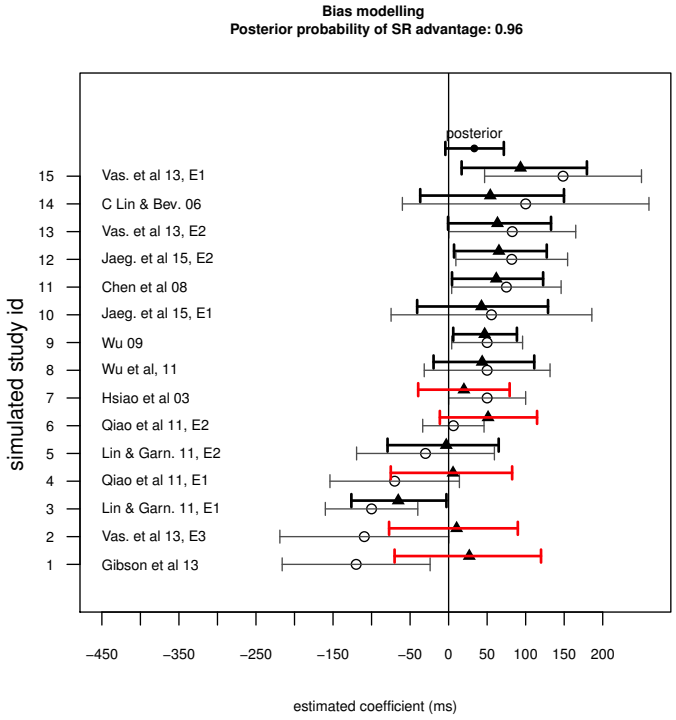


Figure 5.7: Bias modelling results, with biases included in five studies (the posterior credible intervals of the five studies are marked in red). The remaining studies are assumed to have no bias.

the biases, and preferably these experts would be stakeholders (holding different theoretical positions).

5.10 Conclusion

In this chapter, we identified priors for the biases in five studies, and fitted a bias-adjusted model. We first validated the bias-adjusted model, written in JAGS, using simulated data. We also validated the analytical approach of Turner and colleagues by verifying that their formulas can recover the true parameter under repeated sampling of simulated data. Then, we implemented the bias-adjusted random-effects meta-analysis using the available data. Although tentative, our main result is a more clearly positive value for the parameter than we had seen in the standard random-effects meta-analysis, suggesting strong evidence for a subject-relative advantage.

CHAPTER 6

CONCLUDING REMARKS

In this dissertation, we investigated the question of whether, in Mandarin Chinese, object relative clauses are harder to process than subject relative clauses. This is a theoretically important question because the dependency distance hypothesis, instantiated in several current theories of sentence comprehension (Gibson, 2000; Lewis and Vasishth, 2005), makes the surprising prediction that Chinese is unique among the world's languages in having an object relative advantage. If there were evidence for an object-relative advantage in Chinese, this would be a strong confirmation of the dependency distance hypothesis. Unfortunately, the empirical literature on this question has been very unclear; evidence has been found for both the object relative advantage and the subject relative advantage. In order to clarify the facts about Chinese, we first carried out a random-effects meta-analysis. This showed that the data weakly suggest a subject relative advantage. Because the lack of homogeneity across studies could be due to systematic sources of bias, we elicited distributions for different possible biases in five of the studies, and carried out a bias-adjusted meta-analysis. The modelling revealed stronger evidence consistent with the subject-relative advantage. In other words, the claims of the dependency distance account were not supported.

Some important caveats to this conclusion are necessary. The most important issue was that we had access to only one expert for assessing the biases. A comprehensive analysis would require several expert assessors, preferably stakeholders

from both sides of the debate. Another important problem was that the expert did not have clear quantitative insight into possible magnitudes of bias. This is because research in psycholinguistics typically uses null hypothesis significance testing, focussing only on p-values, and ignoring the magnitudes and uncertainty estimates of effects. Therefore, a prerequisite for such bias modelling should be a comprehensive quantitative evaluation of effect sizes across experiments; this will provide a more solid basis for making expert judgements. A third issue is that it is by no means clear that all major biases are being taken into account. A more comprehensive taxonomy of bias types specific to psycholinguistics needs to be developed. A fourth issue is that we have assumed that each bias is independent of the others, and that each study is independent of the others. The independence-of-biases assumption may be more or less defensible, but the independence of studies is questionable; for example, the two experiments by Qiao and colleagues were run by the same researchers, and could be considered to be dependent. Sutton et al. (2012) address this issue by developing several examples of meta-analyses where studies form clusters; in future work, it may be worth extending the analyses presented in this dissertation by making more realistic assumptions about study-level dependencies. Finally, we only analyzed reading times at the head noun, but a full analysis would consider each region of interest separately; this would be a huge undertaking, as expert elicitations are needed for each region of interest separately. In summary, the approach taken here only provides a first approximation of the posterior distribution given data and priors on biases.

Despite these words of caution, it seems clear that such bias modelling is potentially an important tool for synthesizing evidence in psycholinguistics. It may not only clarify the nature of the effects in question, but may also have the side-effect of leading to less biased studies in the future. This could happen if

idealized versions of past studies are repeated); that is, the meta-analysis may also serve to clarify how the researcher can collect better quality data for their own particular research area.

Appendix A

STUDY CHECKLISTS

A.1 Hsiao and Gibson 2003

A.1.1 Internal biases

1. **Selection**:

   (a) Subjects in all conditions recruited from same populations? *This is a within subjects design, so by definition, each subject is his/her control. No improvement is necessary in an idealized design.*

   Effect on study: None expected.

   (b) Subjects recruited over same time periods? *Subjects were recruited over an unspecified period. See page 9 of paper: Forty subjects participated in the experiment. Six were from MIT and the surrounding community. Seven resided in Taiwan, and were at- tending a wedding in California at the time of the experiment. The other 27 were based in and around Los Angeles. All were native speakers of Mandarin Chinese spoken in Taiwan and were naive as to the purposes of the study.*

   An idealized study would do a lab-based experiment with a sample from a more homogeneous population.

   Effect on study: More variance is expected than usual.

   (c) Were inclusion and exclusion criteria clear?

   Yes. The authors state on page 9: Furthermore, although most of the

participants also spoke English, Mandarin Chinese was the primary language that they used in their day-to-day life.

Effect on study: None expected.

(d) Was randomization used?

This is a standard Latin square design. There is no information on how subjects were allocated to each list. Evidence: page 9: "The stimuli were pseudo-randomized separately for each participant so that at least one filler item intervened between two targets." So, randomization of items was used, but it is not clear whether subjects were randomly allocated to each list. One can assume that was not the case, since this was not mentioned.

An idealized study would ensure that allocation to each list is random.

Effect on study: None expected.

(e) Did the comparison conditions constitute a fair comparison (were they minimal pairs)?

Subject and object relatives have different local ambiguities (see Jäger et al. (2015) for a detailed discussion), so that it is not a fair comparison to just compare them without disambiguating the local ambiguities. Specifically, an ambiguity arises at the head noun that may cause a slowdown in SRs.

An idealized study would ensure that all local ambiguities are resolved by the time a head noun is processed.

Effect on study: The observed object relative advantage may be due to the effect of this local ambiguity. Thus, it is possible that the true effect is 0, or even positive in sign.

2. **Performance**:

   (a) Were subjects blinded?

   Yes.

   Effect on study: None expected.

   (b) Was the experimenter blinded?

   No. This is only rarely the case in psycholinguistics.

   Effect on study: None expected.

   (c) Adequate concealment of experimental manipulation (adequate use of filler sentences to mask the experimental manipulation)? *Not clear. An idealized version would have a range of syntactic constructions as fillers, to prevent the subject from detecting that the experiment is about relative clauses.*

   Effect on study: None expected.

   (d) Was the experimental method appropriate?

   Yes.

   Effect on study: None expected.

3. **Attrition**:

   (a) Were any subjects excluded post-hoc?

   In the published paper, data from 35 subjects is reported. In the original analysis (Hsiao's PhD dissertation, data from 32 subjects was reported; the three participants in the original study were removed due to excessively long reading times. These slow subjects were included in the final analysis; they were aged 56, 65, and 69 years. Quote from Hsiao's dissertation: "In addition, three participants' data were excluded from the

analyses due to slow reading times, two standard deviations slower than the mean" (p. 65 of Hsiao dissertation).

An idealized study would not use question-response accuracy as a criterion for exclusion, as this was not standard practice in the other studies.

Effect on study: Removing or including these subjects may have affected the responses. It is difficult to say what effect this could have had on the reported effect.

(b) Are the results likely to be affected by post hoc exclusions?

Not clear, as the original data are unavailable.

Effect on study: Not clear.

4. **Detection**:

(a) Was data analyst blinded?

This is almost never the case in psycholinguistics. An idealized study would blind both the experimenter and the analyst.

Effect on study: None expected.

(b) Reading time measured accurately (appropriate software used, lab conditions)?

No. It appears that the experiment was not conducted in the lab. An idealized study would be done under lab conditions.

Effect on study: The variance is likely to be higher than usual.

(c) Was the statistical analysis appropriate?

Probably not. No model checking seems to have been done, and analysis was on raw reading times, not log-transformed RTs.

Effect on study: The effect may be entirely due to extreme values in one condition, as in Gibson and Wu (2013).

5. **Other**:

   (a) Do you suspect other bias?

   Yes. The effect seen in the paper was driven by almost impossible to comprehend double center embeddings. This drives the main conclusion of the paper. The single embeddings actually show a tendency towards a subject-relative advantage at the head noun. An idealized study would not use double embeddings.

   Effect on study: The double center embedding effect may be due to this problem in the design. But since we ignore double embeddings, we can disregard this problem.

A.1.2 External biases

Population

The Hsiao and Gibson 2003 study deviates from the target population quite substantially. The description of the participants says: "Forty subjects participated in the experiment. Six were from MIT and the surrounding community. Seven resided in Taiwan, and were attending a wedding in California at the time of the experiment. The other 27 were based in and around Los Angeles. All were native speakers of Mandarin Chinese spoken in Taiwan and were naive as to the purposes of the study." This is already a very heterogeneous group compared to standard psycholinguistics studies. Moreover, it appears that the experiment was not done in a laboratory setting. In the published paper, data from 35 subjects is reported. However, in the original analysis (Hsiao's PhD dissertation, data from 32 subjects was reported; the three participants in the study, aged 56, 65, and

69 years, were initially removed due to excessively long reading times. To quote Hsiao's dissertation: "In addition, three participants' data were excluded from the analyses due to slow reading times, two standard deviations slower than the mean". (p. 65 of Hsiao dissertation). These slow subjects were included in the final Hsiao and Gibson 2003 paper; it is not clear what effect these three subjects had on the final reading times, since the data are no longer available.

In an idealized version of this study, the sample could have conformed to the target population better if a more typical homogeneous group of participants (e.g., an undergraduate population) had been used for the experiment. Furthermore, data would not be removed just because of slow responses, since methods exist for stabilizing variance (Box and Cox, 1964).

Effect on study: There are unusually old subjects in this study, and due to the generally slow reading time seen in older populations vs younger baselines, this may cause a slowdown in reading time, leading to larger than usual variance. A further consequence could be that effects may be exaggerated, because we know that longer reading times lead to larger differences between conditions (Wagenmakers and Brown, 2007).

Outcome

The outcome measured was difference in subject vs object relative reading time in single and double center embeddings, at the head noun. The only deviation from the outcome of interest is the use of the double center embedding conditions. However, this should not affect the reading time differences in single embeddings. In order to approximate the idealized version, we will only look at the single center embedding data.

Effect on study: The use of the extremely difficult-to-process double center embeddings may lead subjects to back off to a processing strategy that leads to

their not building full syntactic and semantic representations. A consequence could be that the effect observed at the head noun have a true value of 0 ms, assuming that subjects are not even attempting to complete the dependency between the gap and the head noun. Some evidence for superficial processingis that, among all the studies considered, the lowest question-response accuracies observed are in this study.

A.2 Qiao et al 2011, Expt 1

A.2.1 Internal biases

1. **Selection**:

   (a) Subjects in all conditions recruited from same populations?

   Yes. From pages 5-6: A total of thirty-two native speakers of Mandarin Chinese volunteered to participate in the experiment. Fourteen were current graduate students at the University of Arizona. Eighteen were graduates from the University of Arizona, or the spouses of graduate students. In all cases, Mandarin was their dominant language. This seems like a reasonable pool of subjects.

   Effect on study: None expected.

   (b) Subjects in all conditions recruited from same populations?

   Not clear.

   Effect on study: None expected.

   (c) Were inclusion and exclusion criteria clear?

   Native speakers of Mandarin (no other criteria available).

   Effect on study: None expected.

   (d) Was randomization used?

Not clear. No mention is made of randomisation anywhere in the paper.

Effect on study: None expected.

(e) Did the comparison conditions constitute a fair comparison (were they minimal pairs)?

Yes.

Effect on study: None expected.

2. **Performance**:

(a) Were subjects blinded?

Yes.

Effect on study: None expected.

(b) Was the experimenter blinded?

Presumably not, since the first author is the presumed experimenter.

Effect on study: None expected.

(c) Adequate concealment of experimental manipulation (adequate use of filler sentences to mask the experimental manipulation)?

Yes. Page 6: "In addition a set of 48 filler sentences with varying structures were included in order to prevent participants from preparing for a relative clause sentence."

Effect on study: None expected.

(d) Was the experimental method appropriate?

No. The method requires a conscious grammaticality decision each time a word is encountered. It could be that this interrupts the parsing

process, and could plausibly exaggerate effects. An idealized study would use standard self-paced reading.

Effect on study: It is difficult to judge the direction of the bias. Variance might be higher than usual.

3. **Attrition**:

   (a) Were any subjects excluded post-hoc?

   No.

   Effect on study: None expected.

   (b) Are the results likely to be affected by post hoc exclusions?

   No.

   Effect on study: None expected.

4. **Detection**:

   (a) Was data analyst blinded?

   No.

   Effect on study: None expected.

   (b) Reading time measured accurately (appropriate software used, lab conditions)?

   Yes.

   Effect on study: None expected.

   (c) Was the statistical analysis appropriate?

   No. Extreme values were trimmed; it is not clear what effect the data removal had. An idealized study would not trim data.

Effect on study: The effect may be driven by the data trimming. Fortunately, we have the raw data, so we computed the estimated effect from the data.

5. **Other**:

   (a) Do you suspect other bias?

   Yes. SRs and ORs have local ambiguities which can bias the reading times. An idealized study would have a design that has no such biases.

   Effect on study: The effect may be driven by the biases, as in Experiment 1.

A.2.2 External biases

Population

The population is typical for psycholinguistics.

Effect on study: None expected.

Outcome

The outcome measured was the difference in decision times in SRs vs ORs at the head noun, in the maze task. Its not clear whether the maze task measures processing time in the same way that reading time does. It could plausibly overestimate processing time, since maze task reading times are almost twice as long as self-paced reading times, and it is well-known (Wagenmakers and Brown, 2007) that variance increases linearly with reaction time. An idealized study would run a self-paced reading study in order to make it more comparable to other studies.

Effect on study: The effect may be overestimated.

A.3 Qiao et al 2011, Expt 2

A.3.1 Internal biases

1. **Selection**:

   (a) Subjects in all conditions recruited from same populations?

   Yes. From pages 5-6: A total of thirty-two native speakers of Mandarin Chinese volunteered to participate in the experiment. Fourteen were current graduate students at the University of Arizona. Eighteen were graduates from the University of Arizona, or the spouses of graduate students. In all cases, Mandarin was their dominant language. This seems like a reasonable pool of subjects.

   Effect on study: None expected.

   (b) Subjects in all conditions recruited from same populations?

   Not clear.

   Effect on study: None expected.

   (c) Were inclusion and exclusion criteria clear?

   Native speakers of Mandarin (no other criteria available).

   Effect on study: None expected.

   (d) Was randomization used?

   Not clear. No mention is made of randomisation anywhere in the paper.

   Effect on study: None expected.

   (e) Did the comparison conditions constitute a fair comparison (were they minimal pairs)?

   Yes.

Effect on study: None expected.

2. **Performance**:

   (a) Were subjects blinded?

   Yes.

   Effect on study: None expected.

   (b) Was the experimenter blinded?

   Presumably not, since the first author is the presumed experimenter.

   Effect on study: None expected.

   (c) Adequate concealment of experimental manipulation (adequate use of filler sentences to mask the experimental manipulation)?

   Yes. Page 6: "In addition a set of 48 filler sentences with varying structures were included in order to prevent participants from preparing for a relative clause sentence."

   Effect on study: None expected.

   (d) Was the experimental method appropriate?

   No. The method requires a conscious grammaticality decision each time a word is encountered. It could be that this interrupts the parsing process, and could plausibly exaggerate effects. An idealized study would use standard self-paced reading.

   Effect on study: It is difficult to judge the direction of the bias, but the variance could be higher than usual.

3. **Attrition**:

   (a) Were any subjects excluded post-hoc?

No.

Effect on study: None expected.

(b) Are the results likely to be affected by post hoc exclusions?

No.

Effect on study: None expected.

4. **Detection**:

(a) Was data analyst blinded?

No.

Effect on study: None expected.

(b) Reading time measured accurately (appropriate software used, lab conditions)?

Yes.

Effect on study: None expected.

(c) Was the statistical analysis appropriate?

No. Extreme values were trimmed; it is not clear what effect the data removal had. An idealized study would not trim data.

Effect on study: The effect may be driven by the data trimming. Fortunately, we have the raw data, so we computed the estimated effect from the data.

5. **Other**:

(a) Do you suspect other bias?

Yes. SRs and ORs have local ambiguities which can bias the reading times. An idealized study would have a design that has no such biases.

Effect on study: The effect may be driven by the biases.

A.3.2 External biases

Population

The population is typical for psycholinguistics.

Effect on study: None expected.

Outcome

The outcome measured was the difference in decision times in SRs vs ORs at the head noun, in the maze task. Its not clear whether the maze task measures processing time in the same way that reading time does. It could plausibly overestimate processing time. An idealized study would run a self-paced reading study in order to make it more comparable to other studies.

Effect on study: The effect may be overestimated.

A.4 Gibson and Wu 2013

A.4.1 Internal biases

1. **Selection**:

   (a) Subjects in all conditions recruited from same populations? *This is a within-subjects design, as is standard in psycholinguistics. Thus, by definition, subjects in all conditions come from the same population. No improvement is necessary in an idealized design.*

   Effect on study: None expected.

   (b) Subjects recruited over same time periods?

   This is unclear, but in the absence of any other information, we can assume that the experiment was not done over an extended period.

   Effect on study: None expected.

(c) Were inclusion and exclusion criteria clear? *All participants (ages 18-30) are stated to be native speakers of Mandarin. No improvement is necessary.*

Effect on study: None expected.

(d) Was randomization used? *It is not clear how the list was chosen for each incoming subject. Ideally, each incoming subject should have been assigned to a separate list; not doing this could lead to an over- or underestimate of the effect.*

Effect on study: None expected.

(e) Did the comparison conditions constitute a fair comparison (were they minimal pairs)? *This experiment has the following possible confound. Charles Lin (Effect of thematic order on the comprehension of Chinese relative clauses. Lingua, 140, 180206, 2014) has pointed out that the context sentences in this experiment could have made subject relatives harder to process, since the thematic roles are reversed between the context and target sentence in subject (but not object) relatives. This potential confound is present in the original Gibson and Wu 2013 study as well. An idealized design would have thematic roles appearing in the same order as in the target sentence. The confound is likely to bias the effect to be an overestimate; in fact, the effect could entirely be due to the confound.*

Effect on study: The observed effect (which had a negative sign) could be entirely due to the thematic role reversal; i.e., the true effect could be zero or even have a positive sign.

2. **Performance**:

(a) Were subjects blinded? *Yes.*

Effect on study: None expected.

(b) Was the experimenter blinded? *No.*

Effect on study: None expected.

(c) Adequate concealment of experimental manipulation (adequate use of filler sentences to mask the experimental manipulation)?

Not clear. An idealized version would have a range of syntactic constructions as fillers, to prevent the subject from detecting that the experiment is about relative clauses.

Effect on study: None expected.

(d) Was the experimental method appropriate? *Yes.*

Effect on study: None expected.

3. **Attrition**:

(a) Were any subjects or items excluded post-hoc? *Yes. Three subjects were excluded for having accuracies lower than 70% percent (subjects 10,13,25; accuracies 69, 62, and 69%). One subject, 31, also had an accuracy below 70% but was included in the study; this was because one item (id 12) was also removed from the experiment and that raised this subject's accuray to 73%. Exclusion of subjects based on this criterion is peculiar because in the first paper that showed an object relative advantage, by Hsiao and Gibson (2003), the mean accuracy was approximately 70% and the authors had included subjects despite accuracy below 70%.*

As mentioned above, item 12 was removed from the data. To quote the authors, "Due to a script error, one item (item 12) was not presented

to the participants, leaving 15 items to be analysed." However, this statement seems to be incorrect: We obtained the raw data from Gibson and all item 12 data (except that for subject 27, see below) were available. The presence of this data has a major impact on the final analyses; the difference between RC types is larger (in favour of the object relative advantage) when this item is removed, and the comparison is statistically significant only if this item is removed. In the analysis shown below in Table A.1, we include all subjects (n=40) and fit a linear mixed model on raw reading times using varying intercepts for subjects and items (more complex models did not converge). Models excluding subjects 10, 13, 25 and item 12 increased the t-value from 1.86 to 2.1.

	Item 12 removed (592 observations)		
	Estimate	Std. Error	t value
(Intercept)	475.93	54.22	8.779
Type (OR 0, SR 1)	112.88	44.58	2.532
	Full data-set (632 observations)		
	Estimate	Std. Error	t value
(Intercept)	512.88	55.00	9.325
Type (OR 0, SR 1)	97.73	52.53	1.860

Table A.1: Linear mixed model analyses of the Gibson and Wu 2013 data excluding and including item 12; the analysis had varying intercepts for subjects and items. The published result is the one excluding item 12.

Finally, 2 out of 7 data points from object relatives and 6 out of 8 data points from subject relatives were missing for one subject (27). This missingness is highly unusual in self-paced reading studies and the only plausible scenario where this can happen is when the experiment is

aborted part-way. The published paper does not discuss the reason for the missingness.

Effect on study: Removing these subjects increased the effect size by 9 ms. The missing data from subject 27 may increased the bias. Removal of the three subjects and item 12 leads to a 15 ms increase in the effect, in favour of the object relative advantage.

(b) Are the results likely to be affected by post hoc exclusions? *Yes. The effect is quite dramatically affected by exclusions.*

Effect on study: See above.

4. **Detection**:

(a) Was data analyst blinded? *No; the reanalysis for the present paper was done by Shravan Vasishth.*

Effect on study: This can lead to biases in the analyses (p-value hacking, or garden-of-forking-paths effects, Gelman and Loken (2013)). However, since we ran a pre-determined analysis, no bias is expected.

(b) Reading time measured accurately (appropriate software used, lab conditions)? *No. Subject 27's data was not collected correctly; there was apparently an aborted experiment for this subject, but this is not reported in the paper.*

Effect on study: It is impossible to say how this missing data could influence the result, so we ignore this issue.

(c) Was the statistical analysis appropriate? *The original analysis was not correct, as it ignored checks for model assumptions being satisfied. However, the data were reanalyzed by Shravan Vasishth, so there is no concern.*

Effect on study: None expected.

5. **Other**:

   (a) Do you suspect other bias? *No.*

   Effect on study: None.

A.4.2 External biases

Population

All participants are stated to be native speakers of Mandarin. This matches the target population.

Effect of study: None expected.

Outcome

The reading time at the head noun was measured.

Effect of study: None expected.

Appendix B

R AND JAGS CODE

B.1 Important R code and functions used

B.1.1 Code for generating simulated data in simulation 1

```
theta<- 15 ## true effect
n<-15 ## sample size
tau<- 0.01 ## between study sd

y <-thetai<-rep(NA,n)
se <- 3

thetai <- rnorm(n,mean=theta,sd=tau)

for(i in 1:n){
y[i] <- rnorm(1,mean=thetai[i],sd=se)
}

datfake <- list(y = y,
            s = rep(se,n),
            n = n)
```

B.1.2 Code for generating simulated biased data in simulation 3

```
theta <- 15 ## true effect
n <- 15 ## sample size
tau <- 0.01 ## between study sd
se <- 3
## functions to generate bias mean and sd:
m <- 1    ## mean of bias
v <- 0.5  ## variance of bias

getbias<-function(nsamp=1,m=1,s_bias=sqrt(0.5)){
  return(rnorm(nsamp,mean=m,sd=s_bias))
}
```

```
gendatsim3<-function(ntrials=15,d=15,
            stddev=0.01,biasm=1,biasv=0.5,se=3){
  n<-ntrials
  s<-rep(se,n)
  theta<-d
  tau<-stddev
  m<-biasm
  v<-biasv
  ## create matrices to store bias means and sd's:
  biasmuint<-matrix(rep(NA,n*5),ncol=5)
  biassdint<-matrix(rep(NA,n*5),ncol=5)
  biasmuext<-matrix(rep(NA,n*2),ncol=2)
  biassdext<-matrix(rep(NA,n*2),ncol=2)

  for(i in 1:n){
    biasmuint[i,]<-getbias(nsamp=5)
    biasmuext[i,]<-getbias(nsamp=2)
    biassdint[i,]<-rep(sqrt(v),5)
    biassdext[i,]<-rep(sqrt(v),2)
  }

thetai <- rnorm(n,mean=theta+biasmuext[,1]+biasmuext[,2],
                sd=tau+biassdext[,1]+biassdext[,2])

y<-rnorm(n,mean=thetai+
                biasmuint[,1]+biasmuint[,2]+
                biasmuint[,3]+biasmuint[,4]+
                biasmuint[,5],
                sd=se+biassdint[,1]+biassdint[,2]+
                biassdint[,3]+biassdint[,4]+
                biassdint[,5])

biased_se<-(se+tau+biassdint[,1]+biassdint[,2]+
                     biassdint[,3]+biassdint[,4]+
                     biassdint[,5]+biassdext[,1]+
                     biassdext[,2])

datfake<-list(y=y,
              s=biased_se,
              n=n
              )
return(datfake)
```

```
}
```

B.1.3 Code for generating simulated data in simulation 4

```
theta<- 15 ## true effect
n<- 15 ## sample size
tau<- 0.01 ## between study sd
se <- 3
## functions to generate bias mean and sd:
m<-1    ## mean of bias
v<-0.5  ## variance of bias

getbias<-function(nsamp=1,m=1,s_bias=sqrt(0.5)){
  return(rnorm(nsamp,mean=m,sd=s_bias))
}

gendat4<-function(ntrials=15,d=15,stddev=0.01,
         biasm=1,biasv=0.5,se=3){
  n<-ntrials
  #se<-rnorm(n,mean=70,sd=10)
  s<-rep(se,n)
  theta<-d
  tau<-stddev
  m<-biasm
  v<-biasv
  ## create matrices to store bias means and sd's:
  biasmuint<-matrix(rep(NA,n*5),ncol=5)
  biassdint<-matrix(rep(NA,n*5),ncol=5)
  biasmuext<-matrix(rep(NA,n*2),ncol=2)
  biassdext<-matrix(rep(NA,n*2),ncol=2)

  for(i in 1:n){
    biasmuint[i,]<-getbias(nsamp=5)
    biasmuext[i,]<-getbias(nsamp=2)
    biassdint[i,]<-rep(sqrt(v),5)
    biassdext[i,]<-rep(sqrt(v),2)
  }

thetai <- rnorm(n,mean=theta+biasmuext[,1]+biasmuext[,2],
                sd=tau+biassdext[,1]+biassdext[,2])

y<-rnorm(n,mean=thetai+
                biasmuint[,1]+biasmuint[,2]+
                biasmuint[,3]+biasmuint[,4]+
                biasmuint[,5],
```

```
                sd=se+biassdint[,1]+biassdint[,2]+
                biassdint[,3]+biassdint[,4]+
                biassdint[,5])

biased_se<-(se+tau+biassdint[,1]+biassdint[,2]+
                    biassdint[,3]+biassdint[,4]+
                    biassdint[,5]+biassdext[,1]+
                    biassdext[,2])

    ## here, because there are so many variance components
    ## I convert to precision at the outset:
    datfake<-list(y=y,
                  p=1/(biased_se)^2,
                  n=n,
                  iselmu = biasmuint[,1],
                  iperfmu = biasmuint[,2],
                  iattrmu = biasmuint[,3],
                  idetmu = biasmuint[,4],
                  iothmu = biasmuint[,5],
                  epopmu = biasmuext[,1],
                  eoutmu = biasmuint[,2],
                  iselprec = 1/biassdint[,1]^2,
                  iperfprec = 1/biassdint[,2]^2,
                  iattrprec = 1/biassdint[,3]^2,
                  idetprec = 1/biassdint[,4]^2,
                  iothprec = 1/biassdint[,5]^2,
                  epopprec = 1/biassdext[,1]^2,
                  eoutprec = 1/biassdext[,2]^2)
return(datfake)
}
```

B.2 JAGS code for standard random-effects meta-analysis

B.2.1 Code for random-effects meta-analysis

```
    model
{
    for( i in 1:n )
{
    p[i] <- 1/s[i]^2
    y[i] ~ dnorm(thetai[i],p[i])
    thetai[i] ~ dnorm(theta,prec)
}
    ## priors for theta:
    ## theta lies between (-1.96*100,1.96*100):
    theta ~ dnorm(0,1/100^2)
    tau ~ dunif(0,200)
    tau.sq <- tau*tau
    prec<-1/(tau.sq)
    ##generate posterior predicted values:
    ## (not shown in dissertation)
    pred ~ dnorm(theta,prec)
}
```

B.2.2 Code for bias-adjusted random-effects meta-analysis

```
    model
    {
    for( i in 1:n )
    {
    thetai[i] ~ dnorm(theta+epop[i]+eout[i],prec)
    y[i] ~ dnorm(thetai[i]+isel[i]+
                          iperf[i]+
                          iattr[i]+
                          idet[i]+
                          ioth[i],p[i])
    }
    ## priors for theta:
    ## theta lies between (-1.96*100,1.96*100):
    theta ~ dnorm(0,1/100^2)

    for(i in 1:n){
    ## priors on internal biases:
    isel[i] ~ dnorm(iselmu[i],iselprec[i])
    iperf[i] ~ dnorm(iperfmu[i],iperfprec[i])
    iattr[i] ~ dnorm(iattrmu[i],iattrprec[i])
    idet[i] ~ dnorm(idetmu[i],idetprec[i])
```

```
ioth[i] ~ dnorm(iothmu[i],iothprec[i])
## priors on external biases:
epop[i] ~ dnorm(epopmu[i],epopprec[i])
eout[i] ~ dnorm(eoutmu[i],eoutprec[i])
}
## uniform prior tau:
tau ~ dunif(0,200)
tau.sq <- pow(tau,2)
prec <- 1/tau.sq
}
```

Appendix C

SHELF ELICITATION FORMS

The figures shown below display the SHELF elicitation forms for the five studies considered for bias adjustment. Note that although we present truncated normal distributions as elicited distributions, these were truncated only for elicitation purposes, to allow the expert to anchor his judgements. In the JAGS models, we use untruncated normal distributions as priors for the various bias categories.

ELICITATION RECORD – Part 1 – Context

Elicitation title	Bias Modelling (MSc Dissertation Shravan Vasishth)
Session	Elicitation of five experiments' biases
Date	3 Sept 2015
Part 1 start time	14:40

Attendance and roles	Facilitator: Shravan Vasishth Expert: Shravan Vasishth
Purpose of elicitation	To elicit priors for internal and external biases in a Bayesian random-effects meta-analysis of data on Chinese relative clauses. The quantities to be elicited are the means (in millisecond units) and variances of each bias identified in the bias checklist (see Table 4.1 in dissertation). The elicited priors will be used in a model in order to generate bias-adjusted posteriors of the effect.
This record	Participants are aware that this elicitation will be conducted using the Sheffield Elicitation Framework, and that this document, including attachments, will form a record of the session.
Orientation and training	None needed, but the facilitator+expert has read the book by O'Hagan et al (2006), as preparation for this exercise. O'Hagan, A., Buck, C. E., Daneshkhah, A., Eiser, J. R., Garthwaite, P. H., Jenkinson, D. J., ... & Rakow, T. (2006). *Uncertain judgements: Eliciting experts' probabilities*. John Wiley & Sons.
Participants' expertise	The expert is a psycholinguist, specializing in sentence comprehension research, including the processing of relative clauses.
Declarations of interests	The expert has a potential conflict of interest as he has published papers showing a subject-relative advantage (which translates to a positive value for the parameter of interest). He may be biased towards arguing for a subject-relative advantage. However, his own computational model of sentence processing (Lewis & Vasishth, 2005) predicts an object advantage (a negative value for the parameter).
Strengths and weaknesses	A big weakness of this exercise is that the expert has no quantitative knowledge of effect sizes; this is because it is not normal in psycholinguistics to keep track of the magnitude of the effect. Rather, the sign of the effect is of primary interest. Another potential weakness is possible bias; the expert will endeavour to remain detached from the question.

Figure C.1: Pre-elicitation SHELF form (page 1).

Evidence	The evidence is discussed in the dissertation (chapters 1 and 2).
Structuring	*Not applicable, since we are only eliciting priors on biases regarding a difference in means.*
Definitions	For definitions of each bias, see dissertation (section 4.2). The key quantity of interest is the difference (measured in milliseconds) between subject and object relative clause reading times at the head noun of Chinese relative clauses. For more details on what this means, please see the dissertation (chapter 1).

Part 1 end time	15:12
Attachments	Dissertation

Figure C.2: Pre-elicitation SHELF form (page 2).

ELICITATION RECORD – Part 2 – Distribution

Quartile Method

Elicitation title	Bias Modelling (MSc Dissertation Shravan Vasishth)
Session	Elicitation of biases for Gibson and Wu 2013
Date	4th September 2015
Quantity	The bias on the difference (in milliseconds) between subject and object relative clause reading times, at the head noun.
Start time	16:30

Definition	The key quantity of interest is the bias induced on the difference (measured in milliseconds) between subject and object relative clause reading times at the head noun of Chinese relative clauses. For more details on what this means, please see the dissertation (chapter 1). We will call this quantity B.
Evidence	For each bias type in the bias checklist (see Table 4.2 of Dissertation), we will use the bias checklist as evidence.
Plausible range	**Internal biases** **Selection bias**: 0 to -200 ms is the range of possible values. The thematic role reversal could have a large effect on B. **Performance bias**: no bias expected. **Attrition bias**: 0-50 ms. **Detection bias**: no bias expected. **Other**: no bias expected. **External biases** **Population**: no bias expected. **Comparison bias**: See selection bias.
Median	**Selection**: -110 ms **Attrition**: -26 ms
Upper and lower quartiles	**Selection**: -160, -50 ms **Attrition**: -39, -13 ms
Fitting	
Group elicitation	-
Fitting and feedback	**Selection**: Initially, a lower bound of -200 and upper bound of 0 were recorded as the possible bounds of the selection bias.

Figure C.3: Elicitation form for study 1 (page 1).

	Then, the median of -110 was elicited, and the lower and upper quartiles, -160 and -50. The number -110 was settled on by considering the fact that the thematic role confound would contribute about -90 ms to an object relative advantage, and the exclusion of item 12 (see main text) would result in a contribution of -15 ms to the OR advantage. A density plot as feedback showed a distribution consistent with the expert's belief that the most probable values will be negative. The 0.05^{th} and 0.95^{th} quantiles were -160 and -50. This led to the final distribution, shown below. **Attrition**: Initially, a lower bound of -50 and upper bound of 0 were recorded as the possible bounds of the selection bias. Then, the median of -26 was elicited, and the lower and upper quartiles, -39 and -13. This led to the 0.05^{th} and 0.95^{th} quantiles with values -52 and 0.53. This led to the final distribution, shown below.
Chosen distribution	**Selection**: Normal(-107,64^2) I(-200,0) **Attrition**: Normal(-25.5,15.8^2)I(-50,0)
Discussion	Here, and in the entire elicitation process, it is important to note that in psycholinguistics (and psychology and linguistics in general), it is not normal for a researcher to know what a plausible value could be for a particular effect. Experts in these areas only rely on whether an effect is positive or negative in sign, the magnitude and the uncertainty associated with the effect is not typically tracked. An obvious prerequisite for conducting such an elicitation, one which goes beyond the scope of the dissertation, is to tabulate a large range of known effect sizes from the field, along with their uncertainty estimates. Therefore, the entire elicitation process should be considered very tentative.

End time	17:00
Attachments	

Figure C.4: Elicitation form for study 1 (page 2).

ELICITATION RECORD – Part 2 – Distribution

Quartile Method

Elicitation title	Bias Modelling (MSc Dissertation Shravan Vasishth)
Session	Elicitation of biases for Expt 3 of Vasishth et al 2013
Date	4th September 2015
Quantity	The bias on the difference (in milliseconds) between subject and object relative clause reading times, at the head noun.
Start time	06:15

Definition	The key quantity of interest is the bias induced on the difference (measured in milliseconds) between subject and object relative clause reading times at the head noun of Chinese relative clauses. For more details on what this means, please see the dissertation (chapter 1). We will call this quantity B.
Evidence	For each bias type in the bias checklist (see Table 4.2 of Dissertation), we will use the bias checklist as evidence.
Plausible range	**Internal biases** **Selection bias**: 0 to -200 ms is the range of possible values. The thematic role reversal could have a large effect on B. **Performance bias**: no bias expected. **Attrition bias**: no bias expected. **Detection bias**: no bias expected. **Other**: no bias expected. **External biases** **Population**: no bias expected. **Comparison bias**: See selection bias.
Median	-90 ms
Upper and lower quartiles	-107, -73 ms
Fitting	
Group elicitation	-
Fitting and feedback	Initially, a lower bound of -200 and upper bound of 0 were recorded as the possible bounds of the selection bias. Then, the median of -60 was elicited, and the lower and upper quartiles, -140 and -55. A density plot as feedback showed too much

Figure C.5: Elicitation form for study 2 (page 1).

	probability mass in the positive region, which is inconsistent with the expert's belief that the most probable values will be negative. So, an adjustment was made to the median, changing it to -90, and the 0.05^{th} and 0.95^{th} quantiles such that the revised values were -73 and -47. This led to the final distribution, shown below.
Chosen distribution	Normal(-90,25^2) I(-200,0)
Discussion	Here, and in the entire elicitation process, it is important to note that in psycholinguistics (and psychology and linguistics in general), it is not normal for a researcher to know what a plausible value could be for a particular effect. Experts in these areas only rely on whether an effect is positive or negative in sign, the magnitude and the uncertainty associated with the effect is not typically tracked. An obvious prerequisite for conducting such an elicitation, one which goes beyond the scope of the dissertation, is to tabulate a large range of known effect sizes from the field, along with their uncertainty estimates. Therefore, the entire elicitation process should be considered very tentative.

End time	07:54
Attachments	

Figure C.6: Elicitation form for study 2 (page 2).

ELICITATION RECORD – Part 2 – Distribution

Quartile Method

Elicitation title	Bias Modelling (MSc Dissertation Shravan Vasishth)
Session	Elicitation of biases for Expt 1 of Qiao Et Al 2011
Date	4th September 2015
Quantity	The bias on the difference (in milliseconds) between subject and object relative clause reading times, at the head noun.
Start time	18:10

Definition	The key quantity of interest is the bias induced on the difference (measured in milliseconds) between subject and object relative clause reading times at the head noun of Chinese relative clauses. For more details on what this means, please see the dissertation (chapter 1). We will call this quantity B.
Evidence	For each bias type in the bias checklist (see Table 4.2 of Dissertation), we will use the bias checklist as evidence.
Plausible range	**Internal biases** **Selection bias**: no bias expected. **Performance bias**: no bias expected. **Attrition bias**: no bias expected. **Detection bias**: no bias expected. **Other**: 0 to -100 ms is the range of possible values. The local ambiguity just before the head noun could be responsible for the effect observed. **External biases** **Population**: no bias expected. **Outcome bias**: 0 to -50 ms is the range of possible values. An overall longer reading time might exaggerate B.
Median	**Other**: -50 ms **Outcome**: -25 ms
Upper and lower quartiles	**Other**: -25, -75 ms **Outcome**: 2.8, -53 ms
Fitting	-
Group elicitation	-

Figure C.7: Elicitation form for study 4 (page 1).

Fitting and feedback	**Other**: Initially, a lower bound of -100 and upper bound of 0 were recorded as the possible bounds of the selection bias. Then, the median of -50 was elicited, and the lower and upper quartiles, -75 and -25. The density plot as feedback seemed reasonable. The 0.05^{th} and 0.95^{th} quantiles -100 and 0.96. This led to the final distribution, shown below. **Outcome**: Initially, a lower bound of -50 and upper bound of 0 were recorded as the possible bounds of the selection bias. Then, the median of -25 was elicited, and the lower and upper quartiles, -45 and -10. The density plot as feedback seemed reasonable. The 0.05^{th} and 0.95^{th} quantiles -53 and 2.8. This led to the final distribution, shown below.
Chosen distribution	**Other**: Normal(-50,31^2) I(-100,0) **Outcome**: Normal(-25,17^2) I(-50,0)
Discussion	Here, and in the entire elicitation process, it is important to note that in psycholinguistics (and psychology and linguistics in general), it is not normal for a researcher to know what a plausible value could be for a particular effect. Experts in these areas only rely on whether an effect is positive or negative in sign, the magnitude and the uncertainty associated with the effect is not typically tracked. An obvious prerequisite for conducting such an elicitation, one which goes beyond the scope of the dissertation, is to tabulate a large range of known effect sizes from the field, along with their uncertainty estimates. Therefore, the entire elicitation process should be considered very tentative.

End time	18:45
Attachments	

Figure C.8: Elicitation form for study 4 (page 2).

ELICITATION RECORD – Part 2 – Distribution

Quartile Method

Elicitation title	Bias Modelling (MSc Dissertation Shravan Vasishth)
Session	Elicitation of biases for Expt 2 of Qiao Et Al 2011
Date	5th September 2015
Quantity	The bias on the difference (in milliseconds) between subject and object relative clause reading times, at the head noun.
Start time	10:30

Definition	The key quantity of interest is the bias induced on the difference (measured in milliseconds) between subject and object relative clause reading times at the head noun of Chinese relative clauses. For more details on what this means, please see the dissertation (chapter 1). We will call this quantity B.
Evidence	For each bias type in the bias checklist (see Table 4.2 of Dissertation), we will use the bias checklist as evidence.
Plausible range	**Internal biases** **Selection bias**: no bias expected. **Performance bias**: no bias expected. **Attrition bias**: no bias expected. **Detection bias**: no bias expected. **Other**: 0 to -100 ms is the range of possible values. The local ambiguity just before the head noun could be responsible for the effect observed. **External biases** **Population**: no bias expected. **Outcome bias**: 0 to -100 ms is the range of possible values. An overall longer reading time might exaggerate B.
Median	**Other**: -50 ms **Outcome**: -59 ms
Upper and lower quartiles	**Other**: -25, -75 ms **Outcome**: -110, -0.22 ms
Fitting	-
Group elicitation	-

Figure C.9: Elicitation form for study 6 (page 1).

Fitting and feedback	**Other**: Initially, a lower bound of -100 and upper bound of 0 were recorded as the possible bounds of the selection bias. Then, the median of -50 was elicited, and the lower and upper quartiles, -75 and -25. The density plot as feedback seemed reasonable. The 0.05^{th} and 0.95^{th} quantiles -100 and 0.96. This led to the final distribution, shown below. **Outcome**: Initially, a lower bound of -100 and upper bound of 0 were recorded as the possible bounds of the outcome bias; this bias arises from the L-maze task, which encourages lexical processing and discourages syntactic parsing. We therefore expect that subjects will react more adversely to the non-canonical local word order sequencing (Verb-N-de-N) of subject relatives than the more canonical sequencing (N-V-de-N) in object relatives. Then, the median of -59 was elicited, and the lower and upper quartiles, -83 and -25. The density plot as feedback seemed reasonable. The 0.05^{th} and 0.95^{th} quantiles -110 and -0.22. This led to the final distribution, shown below.
Chosen distribution	**Other**: Normal(-50,31^2) I(-100,0) **Outcome**: Normal(-55.6,33.6^2) I(-100,0)
Discussion	Here, and in the entire elicitation process, it is important to note that in psycholinguistics (and psychology and linguistics in general), it is not normal for a researcher to know what a plausible value could be for a particular effect. Experts in these areas only rely on whether an effect is positive or negative in sign, the magnitude and the uncertainty associated with the effect is not typically tracked. An obvious prerequisite for conducting such an elicitation, one which goes beyond the scope of the dissertation, is to tabulate a large range of known effect sizes from the field, along with their uncertainty estimates. Therefore, the entire elicitation process should be considered very tentative.

End time	10:42
Attachments	

Figure C.10: Elicitation form for study 6 (page 2).

ELICITATION RECORD – Part 2 – Distribution

Quartile Method

Elicitation title	Bias Modelling (MSc Dissertation Shravan Vasishth)
Session	Elicitation of biases for Hsiao and Gibson 2003
Date	5th September 2015
Quantity	The bias on the difference (in milliseconds) between subject and object relative clause reading times, at the head noun.
Start time	11:00

Definition	The key quantity of interest is the bias induced on the difference (measured in milliseconds) between subject and object relative clause reading times at the head noun of Chinese relative clauses. For more details on what this means, please see the dissertation (chapter 1). We will call this quantity B.
Evidence	For each bias type in the bias checklist (see Table 4.2 of Dissertation), we will use the bias checklist as evidence.
Plausible range	**Internal biases** **Selection bias**: 0 to 80 ms is the range of possible values. The subject relative was read faster by about 50 ms, and this could be due to the relative ease of processing a Noun-de-Noun sequence compared to a Verb-de-Noun sequence (local processing due to inattentiveness and/or experiment being run outside a lab). **Performance bias**: no bias expected. **Attrition bias**: no bias expected. **Detection bias**: no bias expected. **Other**: no bias expected. **External biases** **Population**: no bias expected. **Comparison bias**: See selection bias.
Median	38 ms
Upper and lower quartiles	13, 59 ms
Fitting	
Group elicitation	-

Figure C.11: Elicitation form for study 7 (page 1).

Fitting and feedback	Initially, a lower bound of 0 and upper bound of 50 were recorded as the possible bounds of the selection bias. Then, the median of 38 was elicited, and the lower and upper quartiles, -13 and 59. The resulting density plot had the 0.05^{th} and 0.95^{th} quantiles -6.3 and 81. This led to the final distribution, shown below.
Chosen distribution	Normal(37.4,26.5^2) I(0,80)
Discussion	Here, and in the entire elicitation process, it is important to note that in psycholinguistics (and psychology and linguistics in general), it is not normal for a researcher to know what a plausible value could be for a particular effect. Experts in these areas only rely on whether an effect is positive or negative in sign, the magnitude and the uncertainty associated with the effect is not typically tracked. An obvious prerequisite for conducting such an elicitation, one which goes beyond the scope of the dissertation, is to tabulate a large range of known effect sizes from the field, along with their uncertainty estimates. Therefore, the entire elicitation process should be considered very tentative.

End time	11:28
Attachments	

Figure C.12: Elicitation form for study 7 (page 2).

BIBLIOGRAPHY

B. Bartek, R. L. Lewis, S. Vasishth, and M. Smith. In search of on-line locality effects in sentence comprehension. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 37(5):1178–1198, 2011.

D. Bates, M. Maechler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, In Press.

G. E. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 211–252, 1964.

D.-G. D. Chen and K. E. Peace. *Applied Meta-analysis with R.* CRC Press, 2013.

C. Clifton, A. Staub, and K. Rayner. Eye Movements in Reading Words and Sentences. In R. V. Gompel, M. Fisher, W. Murray, and R. L. Hill, editors, *Eye movements: A window on mind and brain*, chapter 15. Elsevier, 2007.

J. Cohen. *Statistical power analysis for the behavioral sciences.* Lawrence Erlbaum, Hillsdale, NJ, 2 edition, 1988.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled clinical trials*, 7(3):177–188, 1986.

S. Duval and R. Tweedie. Trim and fill: A simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, 56(2): 455–463, 2000.

D. M. Eddy, V. Hasselblad, and R. Shachter. An introduction to a Bayesian method for meta-analysis the confidence profile method. *Medical Decision Making*, 10 (1):15–23, 1990.

F. Engelmann, L. A. Jäger, and S. Vasishth. The determinants of retrieval interference in dependency resolution: Review and computational modeling. Manuscript submitted, 2015.

S. L. Frank, T. Trompenaars, and S. Vasishth. Cross-linguistic differences in processing double-embedded relative clauses: Working-memory constraints or language statistics? *Cognitive Science*, page n/a, 2015.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534, 2006.

A. Gelman and J. Carlin. Beyond power calculations assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6): 641–651, 2014.

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007.

A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expeditionor p-hackingand the research hypothesis was posited ahead of time. *Downloaded January*, 30:2014, 2013.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, third edition, 2014.

E. Gibson. Dependency locality theory: A distance-based theory of linguistic complexity. In A. Marantz, Y. Miyashita, and W. O'Neil, editors, *Image, Language, brain: Papers from the First Mind Articulation Project Symposium*. MIT Press, Cambridge, MA, 2000.

E. Gibson and J. Thomas. Memory limitations and structural forgetting: The perception of complex ungrammatical sentences as grammatical. *Language and Cognitive Processes*, 14(3):225–248, 1999.

E. Gibson and H.-H. I. Wu. Processing Chinese relative clauses in context. *Language and Cognitive Processes*, 28(1-2):125–155, 2013.

G. V. Glass. Primary, secondary, and meta-analysis of research. *Educational researcher*, pages 3–8, 1976.

P. C. Gordon, R. Hendrick, and M. Johnson. Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 27(6):1411–1423, 2001.

S. Greenland and K. O'Rourke. On the bias produced by quality scores in meta-analysis, and a hierarchical view of proposed solutions. *Biostatistics*, 2(4):463–471, 2001.

D. Grodner and E. Gibson. Consequences of the serial nature of linguistic input. *Cognitive Science*, 29:261–290, 2005.

J. Higgins and S. Green. *Cochrane Handbook for Systematics Reviews of Interventions*. Wiley-Blackwell, New York, 2008.

F. P.-F. Hsiao and E. Gibson. Processing relative clauses in Chinese. *Cognition*, 90:3–27, 2003.

S. Husain, S. Vasishth, and N. Srinivasan. Strong expectations cancel locality effects: Evidence from Hindi. *PLoS ONE*, 9(7):1–14, 2014.

L. Jäger, Z. Chen, Q. Li, C.-J. C. Lin, and S. Vasishth. The subject-relative advantage in Chinese: Evidence for expectation-based processing. *Journal of Memory and Language*, 79–80:97–120, 2015. doi: 10.1016/j.jml.2014.10.005. URL http://www.ling.uni-potsdam.de/~vasishth/pdfs/JaegerChenLiLinVasishth2015.pdf.

M. A. Just, P. A. Carpenter, and J. D. Woolley. Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General*, 111(2): 228–238, 1982.

J. King and M. A. Just. Individual differences in syntactic processing: The role of working memory. *Journal of memory and language*, 30(5):580–602, 1991.

R. Levy. Expectation-based syntactic comprehension. *Cognition*, 106:1126–1177, 2008.

R. L. Lewis and S. Vasishth. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29:1–45, May 2005.

Y. Lin and S. Garnsey. Animacy and the resolution of temporary ambiguity in relative clause comprehension in mandarin. *Processing and producing head-final structures*, pages 241–275, 2011.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, 2000.

L. P. Moja, E. Telaro, R. D'Amico, I. Moschetti, L. Coe, and A. Liberati. Assessment of methodological quality of primary studies by systematic reviews:

results of the metaquality cross sectional study. *British Medical Journal*, 330 (7499):1053, 2005.

J. E. Oakley and A. O'Hagan. *SHELF: The Sheffield Elicitation Framework (version 2.0)*. School of Mathematics and Statistics, University of Sheffield, University of Sheffield, UK, 2010. URL http://tonyohagan.co.uk/shelf.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting experts' probabilities*. John Wiley & Sons, 2006.

M. Plummer. JAGS version 3.3.0 manual. *International Agency for Research on Cancer. Lyon, France*, 2012.

X. Qiao, L. Shen, and K. Forster. Relative clause processing in Mandarin: Evidence from the maze task. *Language and Cognitive Processes*, 27(4):611–630, 2012.

M. S. Safavi, S. Husain, and S. Vasishth. Locality and expectation in separable Persian complex predicates. Submitted to Frontiers, 2015.

G. Sampson. *Empirical linguistics*. Continuum, London, 2001.

D. J. Spiegelhalter and N. G. Best. Bayesian approaches to multiple sources of evidence and uncertainty in complex cost-effectiveness modelling. *Statistics in medicine*, 22(23):3687–3709, 2003.

A. Staub. Eye movements and processing difficulty in object relative clauses. *Cognition*, 116(1):71–86, 2010.

J. A. Sterne, M. Egger, and G. D. Smith. Investigating and dealing with publication and other biases. *Systematic Reviews in Health Care: Meta-Analysis in Context, Second Edition*, pages 189–208, 2001.

L. Stowe. Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, 1(3):227–245, 1986. ISSN 0169-0965.

A. J. Sutton, N. J. Welton, N. Cooper, K. R. Abrams, and A. Ades. *Evidence synthesis for decision making in healthcare*, volume 132. John Wiley & Sons, 2012.

S. Thompson, U. Ekelund, S. Jebb, A. K. Lindroos, A. Mander, S. Sharp, R. Turner, and D. Wilks. A proposed method of bias adjustment for meta-analyses of published observational studies. *International journal of epidemiology*, 40(3):765–777, 2011.

M. J. Traxler, R. K. Morris, and R. E. Seely. Processing subject and object relative clauses: Evidence from eye movements. *Journal of Memory and Language*, 47 (1), 2002.

R. Turner, D. Spiegelhalter, G. Smith, and S. Thompson. Bias modelling in evidence synthesis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):21–47, 2008.

S. Vasishth. *Working memory in sentence comprehension: Processing Hindi center embeddings*. Garland Press, New York, 2003. Published in the Garland series Outstanding Dissertations in Linguistics, edited by Laurence Horn.

S. Vasishth and R. L. Lewis. Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, 82(4):767–794, 2006.

S. Vasishth, K. Suckow, R. L. Lewis, and S. Kern. Short-term forgetting in sentence comprehension: Crosslinguistic evidence from head-final structures. *Language and Cognitive Processes*, 25(4):533–567, 2011.

S. Vasishth, Z. Chen, Q. Li, and G. Guo. Processing Chinese relative clauses: Evidence for the subject-relative advantage. *PLoS ONE*, 8(10):1–14, 10 2013.

E.-J. Wagenmakers and S. Brown. On the linear relation between the mean and the standard deviation of a response time distribution. *Psychological review*, 114(3):830, 2007.

F. Wu, E. Kaiser, and E. Andersen. Subject Preference, Head Animacy and Lexical Cues: A Corpus Study of Relative Clauses in Chinese. In H. Yamashita, Y. Hirose, and J. Packard, editors, *Processing and producing head-final structures*, Studies in Theoretical Psycholinguistics, pages 173–193. Springer, 2011.