Developing an Experiment: Techniques and Design

Laurie Stowe & Edith Kaan

Revised version,
Summer 2006

Adapted for LIN 4702c/6708c 'Methods in Psycholinguistics'
Based on the Groningen 2002 version

Edith Kaan
kaan@ufl.edu
Linguistics
University of Florida
4131 Turlington Hall
Gainesville, FL 32611
USA

Laurie A. Stowe
L.A.Stowe@let.uu.nl
Linguistics
Rijksuniversiteit Groningen
Groningen, The Netherlands

Contents

Comments Welcome!

We would like to thank Jonneke Brouw, Karlien Dijkstra, Lenka Kerstens, Monique Lamers, Paulien Rijkhoek and Sjoukje van der Wal for proofreading and commenting upon this text. Anneke Steegstra, Hein van Schie and Rienk Withaar have made contributions to text as well. If you still spot typos, obscure passages or any other problems, please let us know, so we can improve the text for future students.

Laurie Stowe
Edith Kaan
Groningen 2002

The present version is based on the 2002 version used at the University of Groningen. Since most of the examples were Dutch sentences, I replaced these examples with English ones. In addition, E-Prime has been included as stimulus presentation software, and references to DMASTR deleted. The (mainly Dutch) appendices are omitted. Finally, the cartoons have been removed for copyright reasons. Comments remain welcome.

Edith Kaan
Gainesville, Summer 2006

Chapter 1 Hypothesis Formation and Testing

In this chapter we will deal with the first steps in designing an experiment: forming a hypothesis and developing a way to test it (Sections 1 and 2). Moreover, when designing your experiment, you must be sure that the results you eventually find cannot be caused by other factors than the ones you are interested in, and you should be aware of yet other hypotheses that can explain the results equally well. (Section 3). Furthermore, you will learn why hypotheses cannot be confirmed on the basis of a null result (Section 4). In Section 5 we will show how predictions can be formulated in terms of main effects and interactions between factors. Section 6 deals with the question of why you need statistics. In Section 7 you will learn why it is important to consider statistics already in the design stage. We will summarize this chapter in Section 8. Along the way, we will introduce some terms that we will be using throughout the rest of this text.

1.1 Hypothesis Formation: Developing a Line of Argument from Question to Conclusion

Let's introduce you to AB. AB is a psycholinguist, especially interested in sentence processing. She wants to test two types of sentence containing the word 'and', to see whether these are processed differently from each other. Examples are given in (1).

(1) a. The hairdresser recognized the grocer and the doctor in the store.
b. The hairdresser recognized the grocer and the doctor spotted the plumber.

And in (1) can either coordinate two noun phrases (*the grocer* and *the doctor*), as in (1a), or two clauses, as in (1b) (*The hairdresser recognized the grocer* and *The doctor spotted the plumber*). AB's question is given in (2)

(2) AB's initial question
Is a NP coordination easier to understand than a clausal (S, CP) coordination?

The first thing that you (and AB), should be asking yourself is: why is this an interesting question? Of course, you can test the difference between NP and S coordination just for the fun of it. However, the aim of all psycholinguistic research is to get to know something about the way humans process language. Unless you make clear how a question such as (2) might lead to a model of sentence processing, questions such as (2) are not interesting, and NSF, NIH or other funding agencies will not provide you with the financial means to investigate it. So first, you have to come up with the underlying ***research question***, which indicates how your more specific question (2) fits into the broader perspective of language processing.

In AB's case, the actual research question is something much more like (3)

(3) AB's research question
How do people interpret (temporarily) ambiguous sentences?

We can see that sentences like (1) are generally at least temporarily ambiguous: the first few words cannot determine which structure the sentence will actually turn out to contain, see (4):

(4) The hairdresser recognized the grocer and the doctor…

Was the doctor recognized by the hairdresser, or did the doctor himself do something? Since readers and listeners are primarily interested in determining the meaning of the sentence, and since ambiguities like (4) are abundant in natural language, the question of how language comprehenders tackle ambiguities is rather important. The answer to this question would give insight in how humans process language.

Once you have formulated the research question, the next step is to formulate one or more ***hypotheses***, i.e. possible answers to the research question, and see in what way you can favor one over the other on the basis of the eventual results of your experiment.

Let us formulate AB's hypotheses. There are several ways in which language comprehenders can tackle temporary ambiguities. One possibility is that they immediately choose only one particular reading when presented with an ambiguity. In choosing they rely on particular strategies (e.g. choose the reading that is syntactically simplest), or the frequency with which particular structures are attested in the language. In the case of the *and*-coordination ambiguities, both factors favor the reading that corresponds to NP coordination. This is AB's first hypothesis:

(5) Hypothesis 1
Given an ambiguity, people choose one reading immediately (on the basis of strategies, frequency…).

Another possibility is that readers/listeners realize that multiple readings are possible, in this case, NP coordination and S-coordination, among others, and that they delay deciding between them until more information becomes available. This is AB's second hypothesis:

(6) Hypothesis 2
Given an ambiguity, people wait until they have disambiguating information to choose one reading over the others.

How can AB turn these two possible answers into something testable? To do that, she has to make some **predictions** that are detailed enough that they can be tested. That is where AB's initial question (2) comes in: Is a NP coordination clause easier to understand than an S-coordination? If people initially choose one possible structure, then we may expect that the other form is harder to understand. So, let us reformulate the first hypothesis, incorporating this prediction:

(7) *Hypothesis 1'*
Given an ambiguity, people choose one reading immediately (on the basis of strategies, frequency…). As a consequence, the chosen reading (in this case, NP-coordination) will be easier to understand than the alternative which is not initially chosen (in this case, the S-coordination).

If, on the other hand, people wait choosing until they have disambiguating information, there will be no difference in difficulty between one reading or the other. Incorporating this prediction, the alternative hypothesis, then will become:

(8) *Hypothesis 2*
Given an ambiguity, people wait until they have disambiguating information to choose one reading over the others. As a consequence, there will be no difference in difficulty between the readings.

However, in order to really test the two hypotheses, we need to make even clearer predictions that specify which pattern of results will be obtained under that hypothesis, and not under Hypothesis 2.

Hypotheses 1 and 2 are actually concerned with what happens at the ambiguity itself. According to Hypothesis 1', at 'and' the reader immediately decides that this is an NP coordination (even though there is no clear disambiguating information). If this is true, readers will be surprised when the sentence continues with a verb, as in *The hairdresser recognized the grocer and the doctor SPOTTED…'*. Therefore, when they read *spotted* in (1b), they try to figure out it how this verb fits in, and possibly revise their analysis of the preceding words. This will lead to an increase in reading times for words that disambiguate towards the non-chosen structure, versus words that are compatible with the structure that was initially chosen. There is no way that Hypothesis 2' would be able to explain this result: if people wait until they have disambiguating information, they can make the correct decision, and there is no surprise or reanalysis. As a consequence, no additional processing time is predicted for one sentence type than the other. AB's final hypotheses, then, are the following:

(9) *Hypothesis 1''*
Given an ambiguity, people choose one reading immediately (on the basis of strategies, frequency…). As a consequence, the chosen reading (in this case, NP coordination) will be easier to understand than the alternative which is not initially chosen (in this case, S-coordination). This will show up as an increase in difficulty when information disambiguates towards the alternative reading (due to reanalysis or a failure to comprehend the structure).

(10) *Hypothesis 2''*
Given an ambiguity, people wait until they have disambiguating information to choose one reading over the others. As a consequence, there will be no difference in processing difficulty between the readings at the point of disambiguation.

In other words, AB can use sentences like those in (1) to test her two hypotheses, because she now has a prediction that is made by one hypothesis, but not by the other, about the processing of these sentence types. If she finds a difference in the predicted direction, then she can say that the results are consistent with Hypothesis 1 but not with Hypothesis 2.

What AB has done is to *develop specific predictions about the behavior to be expected during sentence processing based on the original hypotheses and applied to specific language materials*. Of course, she still needs to choose a method to measure the predicted behavioral difference.

1.2 Hypothesis Testing

The next step, after formulating a research question and hypotheses, is to choose a method of measurement. In Chapter 2, we will consider some methods and their advantages and disadvantages. In the meantime, let us consider AB's experiment. There seem to be two straightforward possibilities, given the form in which the predictions were made. We suggested that people may fail to understand the correct structure for sentences like (1b). If so, AB could simply ask people to assess the grammaticality of the sentence. They ought to think that sentences like (1a) are grammatical significantly more frequently than sentences like (1b). This is a very simple method, since people can simply look through a list of sentences, and circle 'yes' or 'no' at the end of each sentence. Do you see any problems with using this method?

When you try to choose a method, it is usually a good idea to try to imagine that you are a participant carrying out the task. If you are trying your best to carry out this test, you will think about the sentence carefully and eventually make a decision. Additionally, after you have read a number of sentences of this type, you will notice the structures involved (probably). Even if at first you do not recognize that they are grammatical, later in the list of sentences, you probably will. This means that there will not necessarily be enough difference between the two conditions to produce a statistically significant result, even if your prediction is correct. But in order to make the accurate grammaticality judgment, you need to think about the sentences with object relative clauses for a bit longer than sentences like (1a) containing subject relative clauses.

So, instead of simply letting people make a rating ***off-line*** (with all the time in the world to consider their answer), AB can better also measure the time taken to reach the decision. Then, even if there is no difference in the accuracy of the decision, there might be in the ***response time (RT)***. In other words, AB can present the sentences on a computer screen and record the time from presentation of the sentence to when participants make the judgment and look at the RT as well as accuracy of the judgment to see whether there is a difference between NP and S-coordinations. If AB finds that people take more time to respond, or make more errors, to S-coordination than to NP-coordinations, Hypothesis 1'' is confirmed. Or isn't it?

1.3 Further Potential Explanations for Differences between Conditions: When is a Hypothesis TRUE?

Suppose that AB indeed finds a statistically significant difference in the predicted direction between the two conditions. Does this mean Hypothesis 1'' is true? Before answering this question affirmatively, we have to consider whether the difference found between the conditions can be explained on any other grounds than Hypothesis 1''. First, the difference could be caused by (unintended) differences in the materials. Second, even if the difference is due to the experimental manipulation, there probably is an alternative hypothesis that can account for the data as well. So, the strongest conclusion we can draw is that the results **support** Hypothesis 1'', but we **cannot** say that Hypothesis 1'' is the only possible explanation for the results, and hence must be **true**. Let us consider these two points in turn.

1.3.1 Confounding Factors

First, a difference found in the responses to the conditions could be confounded with other factors besides your experimental manipulation. Consider again the sentences in (1), repeated below.

(1) a. The hairdresser recognized the grocer and the doctor in the store.
b. The hairdresser recognized the grocer and the doctor spotted the plumber.

What other differences do we see between the sentences (1a) and (1b) except that the first contains an NP-coordination and the second an S-coordination?

One difference is in the length of the sentences: (1b) has more syllables and letters than (1a). If we assume that a longer sentence takes a bit longer to read, then we could expect a difference in RT between these sentences on the basis of the difference in length just as well as by the difference in structure.

Second, the two sentences contain different words at the sentence end. It is probably true that *in* and *store* are more frequent in English than *spotted* and *plumber*, and hence will read more quickly (See Chapter 3 for how to determine frequency). This is an additional reason why (1b) may be more difficult than (1a).

Another difference between the two sentences is the situation that is expressed. It might be that for some reason it is more plausible that a hairdresser recognizes a doctor in a store than that a doctor spots a plumber. If a sentence is somewhat implausible, people take longer to read it than a more plausible sentence. So, a difference in RT between (1a) and (1b) could also be caused by a difference in plausibility.

This means that, if AB finds that S-coordinations like (1b) take longer to read than NP-coordinations like (1a), this does not mean that her Hypothesis 1'' is correct! The result could have been an ***artifact***, i.e. an unintended result caused by (uninteresting) differences, such as length and plausibility differences, between the conditions.

One solution for AB is to construct materials in which the a and b sentences are closely matched in terms of the length and frequency of the words used. In addition, AB should test the plausibility of the situations expressed (***pretest***, see Chapter 3) and construct her materials such that the two conditions are equally plausible.

Yet another way out is to sentences and predictions a little. One option would be to test sentences such as those in (11).

(11) a. The hairdresser recognized the grocer and the doctor spotted the plumber.
b. The hairdresser recognized the grocer, and the doctor spotted the plumber.

Note that both sentences are S-coordinations. The only difference that (11b) has a comma after *grocer*, which more or less excludes an NP-coordination. Now the words are identical so there is no potential length confound anymore. Additionally, the situation expressed is the same in both versions of the sentence. Under Hypothesis 1, readers will interpret 'and' as an NP-coordination in (12a), but not in (12b). This will lead to an increase in processing difficulty at the verb in (12a) versus (12b), because the verb is incompatible with the reading chosen. The drawback of this design, however, is that we will no longer obtain data for a disambiguated NP-coordination.

The point should be clear: only differences between your experimental conditions should be ones that you are interested in. It is a real nuisance to decide after collecting data that there are other differences, so in the first stage of designing an experiment you have to consider other potential differences between conditions, and take precautions to avoid such artefactual results. The possibility of confounding factors is also why you should *always report your materials in your paper*. That way, other researchers can evaluate them for other potential explanations of the effects that you have found.

1.3.2 Alternative Hypotheses

Let us now turn to the issue of alternative hypotheses. Suppose AB managed to exactly match her conditions (1a) and (1b) in terms of frequency, length and plausibility. Suppose she indeed finds that it takes people longer to process S-coordinations than NP-coordinations. She then concludes that Hypothesis 1 is true: apparently, people do not wait in resolving ambiguities until disambiguating information comes along, but choose one reading right away, leading to an increase in RTs at the point of disambiguation for S-coordinations.

AB reports her results and conclusion in a paper and sends it to a journal. It is reviewed by her greatest rival, CD, who advises the editor to reject it. AB has drawn too strong a conclusion, he thinks, because there is a version of Hypothesis 2 that also accounts for the results. Recall that, according to Hypothesis 2, people wait in assigning a reading to a temporarily ambiguous sentence until they encounter disambiguating information. CD argues that readers might find it more difficult to construct some structures than others. If they do not immediately choose a structure when it is ambiguous, but construct the actual structure at the point of disambiguation, the difference in difficulty will show up at the point of

disambiguation. According to some linguistic theories, S-coordinations are more complex than NP-coordinations. Hence, at the point of disambiguation, object relative clauses will take longer to respond to than subject relative clauses. So AB's Hypothesis 1'' is not the only explanation for the results: the difference between the conditions can also be accounted for under an alternative version of Hypothesis 2.

AB, therefore, cannot conclude that Hypothesis 1'' is true. However, she can conclude something else, namely that Hypothesis 2'' as it is stated in (11) is false: the difference she found between subject and object relative clauses is in no way compatible with a hypothesis that predicts no difference whatsoever between the conditions. The results of her experiment thus forced a revision or greater precision into the formulation of both hypotheses. That in itself is one small step forward for science and one publication for a scientist.

1.4 The Null Result Problem: When is a Hypothesis False?

Now, let us suppose that AB found **no** significant difference between the two conditions. Does this mean Hypothesis 2'' can be favored over Hypothesis 1''? Unfortunately, no. There are numbers of reasons why one could fail to find a statistically significant difference. To mention the two most likely, the measurement technique could have been too insensitive (cf. Chapter 2), or your materials and participants could have differed too much among each other to find a statistically significant difference (we will return to this in Section 7 below). These are alternative explanations that can never be distinguished from Hypothesis 2''. How do you know that the failure to find a difference is due to the workings of the parser and not to your experimental technique? Of course, you could do the same task using all the techniques that are presently available. But still, even if you consistently obtained null-results, you can never exclude that a difference will be detected by a future, more sensitive technique. So, a failure to find a difference can be explained by alternatives to Hypothesis 2''. We therefore cannot say that Hypothesis 2'' is true. All we can conclude is that Hypothesis 2'' is consistent with the data.

Briefly put, a ***null result***, that is no significant difference between the conditions, therefore does not tell you anything. So, to find support for a hypothesis, we have to show that a predicted **difference** exists. To really test Hypothesis 2'', we should think of an experiment where Hypothesis 2'' predicts a difference (of course, not the same difference as Hypothesis 1''). In the next section, we will show how predicted differences can be formulated clearly in terms of main effects and interactions.

1.5 Factors, Main Effects and Interactions

Every experiment has a particular form or ***design***. Roughly put, the design specifies what is compared to what. Let us go through AB's experiment again, while introducing some terminology.

The design of AB's experiment is fairly straightforward. There are two experimental conditions: NP coordination and S-coordination. In an ideal situation, these conditions differ

from each other in only one respect: the kind of coordination. This is the ***explanatory variable*** or ***factor*** in AB's experiment. This factor has two ***levels***: NP-coordination, `+NP' for short; and S-coordination, `+S' for short. Schematically, the design of her experiment looks like this:

(12)

factor: type of coordination	
Level1: +NP	level 2: +S
(1a)	(1b)

Remember AB's hypotheses, repeated here for convenience:

(9) *Hypothesis 1''*
Given an ambiguity, people choose one reading immediately (on the basis of strategies, frequency…). As a consequence, the chosen reading (in this case, NP coordination) will be easier to understand than the alternative which is not initially chosen (in this case, S-coordination). This will show up as an increase in difficulty when information disambiguates towards the alternative reading (due to reanalysis or a failure to comprehend the structure).

(10) *Hypothesis 2''*
Given an ambiguity, people wait until they have disambiguating information to choose one reading over the others. As a consequence, there will be no difference in processing difficulty between the readings at the point of disambiguation.

Hypothesis 1'' predicts a difference between the two levels: the RT to NP-coordination sentences will be different from the RT to S-coordinations, with NP-coordinations being reacted to faster. In other words, a ***main effect*** of the factor Type of Coordination is predicted.

Let us now turn to a more complex design. Suppose AB is interested in an additional question, namely whether ambiguity resolution is influenced by semantic information, in particular, animacy information. She suspects that people tend to assume that an NP is more likely to be a subject it denotes an animate entity, and an object if it denotes an inanimate entity. That is, she has a different version of Hypothesis 1:

(13) *Hypothesis 1'''*
Given a temporary ambiguity such as that found in coordinations, people choose one reading immediately.

They make this choice on the basis of semantic information, such as that an animate noun is likely to be an agent (therefore subject) but an inanimate noun is likely to be a theme or patient (therefore object).

As a consequence, the initially chosen reading (S-coordinations when the NP following 'and' is an animate nouns; NP-coordinations when the NP following 'and'

is an inanimate noun) will be easier to understand than the alternative which is not initially chosen (NP-coordination for animates, S-coordination for inanimates). This difficulty will show up as additional time needed for reanalysis at the point where disambiguating information comes in or as a failure to comprehend the non-preferred structure.

Notice that the predictions of this hypothesis are not the same as those of Hypothesis 1'', although they both start from the claim that people make an immediate decision. This sort of difference is why a hypothesis has to be detailed; very different predictions can follow from a vague hypothesis, depending on what other assumptions are made by the researcher.

To test Hypothesis 1''', AB constructs the following 4 conditions.

(14) a. The hairdresser recognized the brush and the comb at first glance.
b. The hairdresser recognized the brush and the comb was also familiar.
c. The hairdresser recognized the brush and the doctor at first glance.
d. The hairdresser recognized the brush and the doctor was also familiar.

(14a) and (14c) contain NP coordinations; (14b) and (14d) are S-coordinations. One factor in this design is therefore again the Type of Coordination. (14a) and (14b) differ from (14c) and (14d) in terms of the animacy of the noun phrase following 'and'. AB has therefore introduced a second factor in her design. What is this factor? Note that the conditions differ not only concerning the type of relative clause, but also whether the sentence contains an second inanimate NP or not. Let's call this second factor Animacy, with the levels Animate (18c,d) and Inanimate (18a, b).

What are AB's predictions? Up to and including the NP following 'and' the sentences in (14) are ambiguous: they can either be an NP or S-coordination. On the basis of Hypothesis 1''', AB expects that in (14a) and (b) readers opt for an NP-coordination, because the NP following 'and' is inanimate and therefore a good object and bad subject. As a result, reading times will be longer for sentences like (14b) than for sentences like (14a), since the decision turns out to be incorrect and must be revised. In (14c) and (14d) on the other hand, AB expects that the reader assumes that the animate noun phrase *the doctor* is a subject of a new clause. This decision turns out to be incorrect in (14c), so that reading times will be longer for (14c) than for (14d).

A design like the one in (14) with two or more factors is called a **factorial design**. One advantage of such designs is that they help the experimental control for unintended differences between the conditions. In this case, we expect the RTs for (14a) to be shorter than for (14b), and the RTs for (14c) to be longer than for (14d). Note that in both comparisons, we are comparing sentences containing *at first glance* to *was also familiar*. Any differences we find between the two comparisons (14a vs b; 14c vs d), therefore cannot be due to these lexical differences between the sentences -- although such differences still may introduce some noise in the data, see Section 1.7.

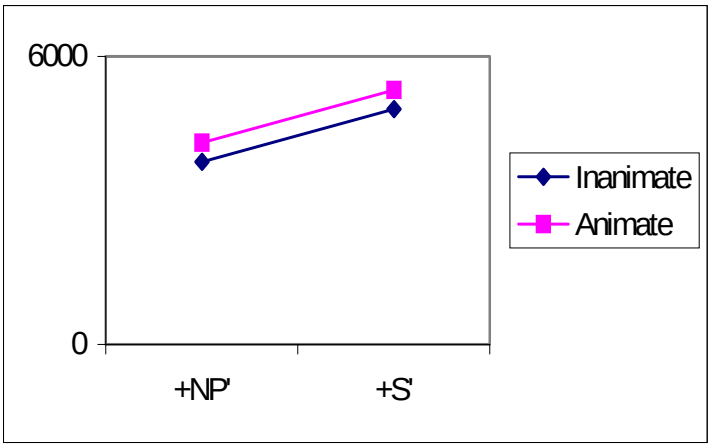
Coming back to AB's predictions, note that the prediction now states that the difference between NP and S coordinated is dependent on the animacy of the noun phrase following 'and'. In other words, AB predicts an ***interaction*** between the two factors in her experiment, that is, a *Type of Coordination x Animacy interaction*. To make this clearer, let us put AB's new design into a schema.

(20)

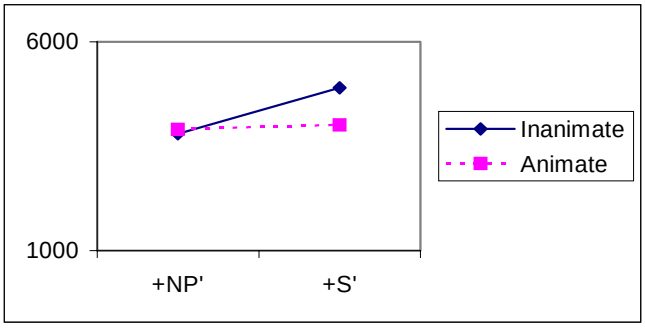
		Coordination type	
		+NP	+S
Animacy	Inanimate	(14a)	(14b)
	Animate	(14c)	(14d)

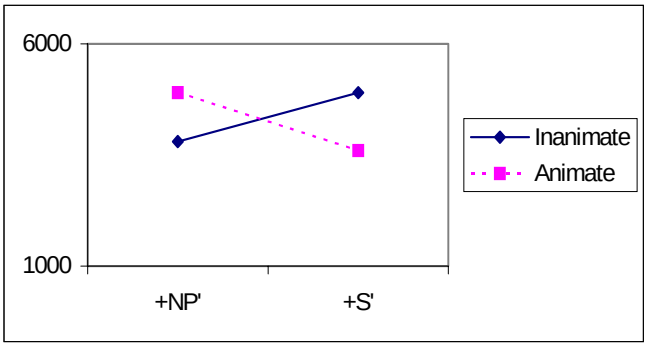
AB's new experiment has two factors: Coordination Type and Animacy, with two levels each. Instead of, or in addition to, predicting a difference between +NP and +S, the main effect of coordination type discussed earlier, AB now predicts that under different conditions of semantic information, the effect of coordination type will turn out to be different. In other words: the difference between +NP and +S in the Inanimate conditions (14a) and (14b) will be different (i.e. larger) than the difference between +NP and +S coordinations in the Animate conditions (14c) and (14d). This (potential) pattern of results is referred to as a Coordination Type x Animacy interaction.

Let us illustrate the terms ***main effect*** and ***interaction*** with some graphs. The figures below display several possible outcomes of AB's experiment. In this experiment, she had people read the entire sentence and press a button when they were done. The first plot represents the results when only main effects are obtained, but no interactions. Reaction times (given in msecs) for the Animate conditions are longer than for the Inanimate conditions, possibly due to differences in plausibility. Reaction times are longer for S coordinations than for NP coordinations, possibly due to the behavior predicted by Hypothesis 1''. In other words, there are main effects of the factor Animacy and of the factor Coordination Type. Note that the difference between NP and S-coordinations is the same regardless of Animacy, hence there is no interaction between the factors.



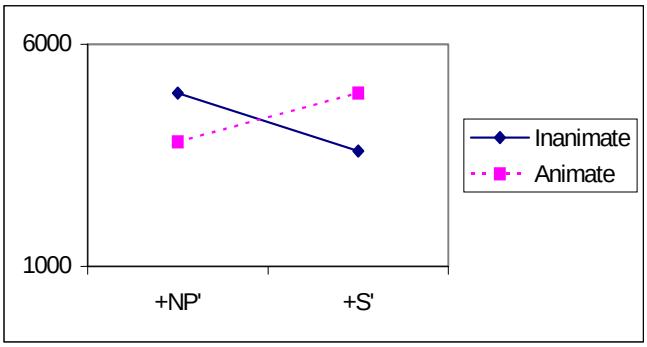
In the figures below, the factors Animacy and Coordination Type do interact: Here, the difference between NP and S-coordinations varies depending on whether the sentence contains an animate or inanimate NP after 'and'. In the Inanimate cases, NP-coordinations are responded to faster than S-coordinations. In the Animate cases, the reaction to S coordinations is as fast as, or faster than NP-coordinations.





However, getting a statistically significant interaction is not enough to support a given hypothesis. The figure below also shows an interaction (the effect of animacy is different depending on whether the sentence contains an NP or S-Coordination). However the pattern

of interaction is not that predicted by the hypothesis discussed above. This is why it is always necessary to look at the actual pattern of responses.



In sum, a hypothesis may predict either a main effect, like Hypothesis 1'' or an interaction between factors like Hypothesis 1'''. When two or more factors are included in the design, interactions must be tested as well as main effects. Interactions are also convenient in those cases where some conditions do not allow direct comparison due to the existence of more than one difference between the conditions: the effect can then be tested indirectly by looking at the difference between differences, as in the comparison of (14a) and (14b) vs. (14c) and (14d). In both cases, what you're trying to obtain as the result of your experiment are *differences*. Differences must be statistically tested. In the following section you will learn why.

1.6 Why We Need Statistics (1): Generalizing Across Participants and Items

AB could let one participant, TS, read her four sentences (14a) through (d) and record this participant's reading times. Suppose TS's reading times are the following.

Sentence			RT in msec
(14a)	: +NP;	inanimate	1769
(14b)	:+S;	inanimate	2598
(14c)	: +NP;	animate	2089
(14d)	: +S;	animate	1990

Clearly, TS took more time to read the S-coordination in (14b) than the NP-coordination in (a), and the difference in TS's reading times between (c) and (d) is smaller than the difference between (a) and (b). So, are AB's predictions (main effect of relative clause type, interaction of Coordination Type by Animacy) borne out?

AB's predictions do hold for participant TS and the four sentences (14a) through (d). However, AB's predictions and hypotheses do not only concern TS and these four sentences; they concern people in general and (ambiguous) sentences in general! What AB would like to

claim is that other participants than TS will show the same pattern of results as well, and that this pattern can also be seen if other sentences than (a) through (d) are used. But she cannot conclude this on the basis of the four reaction times above: The responses might have been coincidentally patterned the way AB expected them to be, and might have been totally different if TS had read other sentences, or if GH instead of TS had read them.

Time, money and mortality will not allow AB to test all language users in the world and all ambiguous sentences that can be constructed in every human language in order to see whether her hypothesis is correct. Instead, she has to select some participants (e.g. 48 adult native speakers of American English) and select some sentences (e.g. 32 sentences of the format (14a) through (d)), conduct the experiment with these participants and sentences, and, on the basis of the results thus obtained, draw some conclusions about people processing sentences in general and about sentences being processed by people in general. Whether AB's results allow her to infer something about a group of people and sentences larger than her sample (or: whether your results can be ***generalized over participants and items***) can be determined by statistically testing the differences found between the conditions.

At this point, we must make one point: hypotheses almost never concern `people in general' but are restricted to subgroups of people. Although this is often not stated explicitly, most psycholinguistic research concerns adult native speakers (undergraduates??!) only.
It should be stated as part of your hypothesis to which group of participants you wish to generalize your results, as this has implications for the kind of participants you select for your experiment. To give a somewhat exaggerated example, if you are interested in language acquisition, and would like to say something about language processing by preschool children, you would probably not choose only college freshmen as your participants. Your participants should always be a representative sample of the more general group of participants to which you wish to generalize your results.

In the next section, we will discuss the basic idea of a statistical test, and show you need to be concerned about statistics already in the design stage of the experiment. We will return to statistical processing in more detail in later chapters.

1.7 Why We Need Statistics (2): Statistically Significant Differences and Noise

A statistical test is used to tell you how likely it is that any difference you find in your experiments is 'real', that is, holds for the group of all possible participants and/or items. Or, to be more precise, a statistical test tells you what the chance is that the difference you find in your experiment is accidental, or not real. The smaller this probability, the more likely that the difference is genuine. If this chance is smaller than .05 (5%), we say that the difference is ***significant***. The chance value is determined on the basis of three things: (i) the means of the responses in each condition, (ii) the spread (variability) of the responses around these means (as expressed by the standard deviation or variance), and (iii) the number of participants and/or sentences, which together determine the number of data points that go into the test.

It should be clear that the more participants and items that you test, the better the estimate of the real difference between conditions. The more data points you have, the better the estimate

of the mean and variability. To illustrate this, suppose you have a barrel of M&Ms. You want to know the mean weight and variability in weight of M&Ms. You will get a better idea of the mean weight if you measure one hundred M&Ms than if you measure only two. The same holds for participants and sentences in an experiment: the more you have, the better idea you get of the 'real' mean and variability in each condition. So, if there is a real but small difference between conditions, you are more likely to detect it using a large number of participants and sentences than when you are only testing a few. This is known as ***statistical power***.

As you will also sense intuitively, the larger the difference in means between two conditions, and the more tightly the responses in each condition are clustered around the mean, the smaller the chance that the two conditions do not differ in reality. For instance, Condition A has a mean (average) RT of 1697 msec, with the fastest response being 1625, and the slowest 1702 msec. This is tested against Condition B with a mean of 2017 and responses ranging from 2001 to 2067 msec. Note that there is no overlap at all in the response times. Under these conditions, the difference between the two conditions will most likely turn out to be significant. On the other hand, if the difference between the means is very small (say, 1697 msec for Condition A vs 1702 msec for Condition B), or the results are very widely spread around the means (say, 1367-2378 msec for Condition A and 1498-2423 msec for Condition B), the chance that the two conditions do not actually differ from each otherwill be fairly high, despite the difference in means. The relationship of the difference between means (or: ***effect size***) and the variability around the mean in each condition is known as the ***signal to noise ratio***.

Now, why do these issues concern us at the design stage of the experiment? We have just seen that the variability (standard deviation, variance) of the response is one of the factors that determine whether a difference is significant. If participants respond to each sentence within an experimental condition in a different way, the variance of the responses within the conditions is large, and chances are low that a difference between the two conditions will be significant. On the other hand, if there is very little variability within conditions, even a relatively small difference between two conditions can be significant. This means that the variability in the responses must be kept as small as possible: the smaller the variance, the larger the likelihood that the differences you are interested in will turn out to be significant.

To give just one more example, suppose that two conditions, A and B, are tested in a reading time experiment. The experimenter's hypothesis is that ambiguous sentences will be more difficult to process than unambiguous sentences. Condition B contains ambiguities; Condition A contains only unambiguous sentences. The experimenter's prediction is that the sentences in Condition B will be reacted to more slowly than the sentences in Condition A. One of the 16 test sentences in Condition A contains the word *archaeopteryx*. This is a word most participants, except for a few dinosaur experts, will not be familiar with. If participants are not familiar with a word, they take longer to respond. As a result, the variance will be larger for Condition A than for Condition B. Let's assume that the 15 remaining sentences in Condition A and all sentences in Condition B only contain familiar words. Not including responses to the *archaeopteryx* sentence, responses will fall within a range of, say, 1300 - 1500 msec for Condition A and 1500-1700 msec for Condition B. However, because of the

unfamiliar word in Condition A, the responses for all sentences in that condition, including the *archaeopteryx* sentence, will range from 1300 to, say, 1800 msec. So, the variance of the responses in Condition A is quite large.

Suppose that the experimenter runs a statistical test on these results and obtains a null result, that is, the difference between the Conditions A and B is not significant. Does this mean that the experimenter's hypothesis that ambiguous sentences (Condition B) take longer to read than unambiguous sentences (Condition A) is wrong? No! A real difference between the two conditions might have been obscured by the large variance of the reading times for the sentences in Condition A. The range of reading times for A could be too large to obtain a significant difference between the two conditions. The use of an unfamiliar word introduced some ***noise*** in the data, which prevented results from being significantly different.

In addition to an increased variance, the inclusion of *archaeopteryx* will result in longer mean reaction times for Condition A than would be the case if no such infrequent word was used. Recall that the experimenter predicted that reaction times would be longer for B than for A. The unfamiliar word in Condition A thus biases the results in the wrong direction, obscuring a potentially real difference between the two conditions.

Although we are concerned here mainly with the issue of noise in the data, it is equally important to note that unfamiliar words in one condition could lead to an artefactual result. Suppose that B instead of A included *archaeopteryx*, and that a significant difference between A and B were obtained, with B being reacted to more slowly than A. If B was predicted to be more difficult than A, one may conclude that the prediction is supported. However, this result may not have been due to the factor of interest (e.g. ambiguity), but may instead have been caused by the use of an unfamiliar word in one condition, which may have biased the reaction times into the predicted direction. (cf. section 1.3).

In sum, one of the things you clearly want to avoid in your experiment is noise caused by a large variability of the responses. Noise will decrease the chance of finding a significant difference between the conditions. To reduce noise as much as possible, you have to construct your materials in an as homogenous way as possible. This is achieved by carefully matching your materials between and within conditions. We will deal with matching in Chapter 3.

Note that noise can also be caused by the participants that participate in your experiment. If you only include students between the age of 18 and 26 in your participant group, the variance will be much smaller than if your participant group was a mixture of children, teenagers and middle-aged people having various educational backgrounds (Chapter 4).

1.8 Summary

The first step in conducting an experiment is to formulate a research question. From this general research question, one or more hypotheses, that is, potential answers to the research question, are derived. Alternative hypotheses can only be tested if they lead to different predictions.

In the design stage of the experiment, you have to indicate precisely what pattern of results (that is, what pattern of differences) is expected under a particular hypothesis, given the method of measurement and the format of the materials you will be using. The clearest way to formulate these predictions is in terms of main effects and interactions between factors.

Before going ahead with the experiment, you must be sure that the results you will eventually find cannot be explained by unintended differences between the conditions (in other words, that your results are not artefactual), and be aware of yet other hypotheses which can explain the results obtained in the experiment.

You must also take care that actual differences between the conditions will not be obscured by too much noise in your data. Too much variability in your materials and participants will increase the variance and might lead to a null result. A null result can also be due to your method of measurement not being sensitive enough. In the next chapter, we will give an overview of some experimental techniques and their sensitivity, among other things.

Chapter 2 Experimental Techniques for sentence processing

It is important to choose the most suitable measure for the hypotheses that you wish to test. For some hypotheses, this may be an extremely simple technique, for others it may be exceptionally complicated. More complicated is not always better. For example, if you wish to test the prediction that "Object relative clauses are more difficult to understand than subject relative clauses", it is useless to put people into a PET scanner: a PET scanner measures brain metabolism, but it does not say anything directly about difficulty in processing. Asking people to decide whether the sentence was grammatical does provide information, although it is a much simpler technique: if the sentences are difficult enough to understand, people will say 'No' more frequently with the difficult sentences than with the easy control sentences.

What are the important issues in choosing a method?

- ***specificity***: is the measurement technique able to measure what you need to in order to answer your question?
- ***sensitivity***: does the task show up reliable differences for the kind of materials you want to use?
- ***controllability***: is it possible to match materials adequately for the method to be used in the particular experiment that you want to carry out?
- ***availability***: do you have access to the method? If not, it is not a very practical method for your experiment, even if it is otherwise the best choice.

By **specificity**, we mean that each task measures a certain variable. For PET, that is brain metabolism; for response time (RT) measures, how long it takes to carry out a certain task. What do you need to know to test your hypotheses? If your prediction concerns brain metabolism, then an RT measure won't serve your purpose. If you want the time taken to carry out a certain task, then PET won't help at all, but an RT task will. Sometimes a task that measures something that you are not directly interested in can nevertheless help to answer your question. It is all a matter of whether there are indirect as well as direct predictions made by your hypotheses (cf. lexical decision below). **You always have to present a clear argument as to why a task is an appropriate measure for the prediction made by your hypothesis**.

Sensitivity is a different issue. It cannot be determined entirely on the basis of logic. The real question is how big a difference do you expect and how small a difference can you measure? The only way to really decide this is to go ahead and do the experiment. If you get a result, then it was sensitive enough (at least if you have managed to avoid all artifacts; see Chapter 3). The obvious question to ask is whether the task has frequently been used in the literature and whether it produced results. If so, you should be able to use it as well.

Sensitivity is also related to the number of participants and items that must be run to obtain a reliable difference. If you want to know the answer to this question, look at other studes using the same technique. You should be able to get a result with comparable numbers.

A third issue is what we will call **controllability**. There are certain factors which affect the way in which any given task is carried out. If you browse the literature, you will find that authors mention certain factors that were matched, since otherwise they cause noise in the data (see Chapter 3). Some tasks are more sensitive than other tasks to certain aspects of the stimuli. For instance, brain potentials (ERPs) are very sensitive to physical characteristics of a word (i.e. luminance of the screen), whereas luminance generally does not affect reading times. So, if you wish to conduct an ERP experiment, differences in luminance between the conditions must be controlled for. If you know that a task is sensitive to a factor, but you are not able to control for that factor, then the task is not suitable for use in your experiment, even if it would otherwise be very appropriate.

In this chapter, we will consider some of the more commonly used experimental techniques for *sentence processing* research, discussing their specificity, sensitivity and controllability. We divided the various techniques into five groups: accuracy measures (Section 1), rating tasks (Section 2), response time measures (Section 3), indirect and concurrent tasks (Section 4) and physiological measures (Section 5). Bear in mind, though, that these distinctions are mainly made for expository reasons rather than being a meaningful classification of methods. We also have no intention of providing an exhaustive overview.

2.1 Accuracy Measures

Accuracy at various judgment tasks is a rather obvious way to tap comprehension. We will discuss four such tasks below.

2.1.1 Sentence Verification

A commonly used accuracy task is picture/sentence verification. In this task, people are asked to judge whether a picture and sentence match.

Specificity
What is picture/sentence verification good for? It reflects the difficulty which people have in understanding sentences. If people have difficulty understanding a sentence, they will also tend to make the verification decision incorrectly.

A classic use of the picture verification task is Dan Slobin's test of children's comprehension of passive sentences in English. He showed that children were better able to understand passives when guided by semantic information.

(1) a. The cake was eaten by the child.
b. The cat was bitten by the dog.

Children were more accurate in verification with sentences like (1a), because they had semantic information to help interpret the structure correctly.

Sensitivity
This task is not very sensitive. The sentence types must be extremely different in complexity before a different in accuracy will show up. The task is most frequently used with children and aphasics, since they are likely to show greater differences between levels of sentence difficulty. If the difference is large, only a few sentences in each condition are necessary; but if it is small, you will need to use a very large number of items to catch people making mistakes at all. On the other hand, you do not need very many participants, if the task works.

Controllability
Controlling possible differences between conditions is also not a simple job. The sentences must not only have comparable lengths, etc., but the pictures must be comparable, too. This leads to the problem of how to match the difficulty of interpreting a picture. Are the participants named in various sentences equally identifiable in your pictures? Are the activities equally identifiable? How do you know? About the only way to succeed in matching the complexity of pictures is to use the same picture for all the conditions you are testing. Is this possible? Only if the actions and the participants can be the same in all sentence conditions, obviously. Using the same pictures can also lead to repetition effects, which may cause difficulties in the design (see Chapter 3 for effects of repetition and some methods of dealing with them).

Some investigators have conducted this task using one sentence to one picture. Frequently however, experimenters use several distractor pictures, since they are interested not only in accuracy but also in the type of error which is made. If you want to draw conclusions from the kind of errors made by your participants, then you need to consider your distractor items **very** carefully (cf. Ch. Koster, 1993).

2.1.2 Comprehension questions

Another accuracy measure is to present text and ask people to answer questions about it. Across the experiments, these questions can be of several types: Yes/No questions, questions where two or more alternative answers are provided or true/false statements.

Specificity
What can be measured with this task? First, you can measure how people understood the text. For instance, one can use comprehension questions to investigate how people interpret globally ambiguous sentences. Second, you can measure the difficulty of understanding the question itself, which can also be varied.

Sensitivity
This is not a very sensitive task, for the same reasons as picture verification. Normal adult participants simply do not make that many mistakes. As with picture verification, the number of items that are necessary is dependent on how many errors people make. If you

wish to use this technique, you might consider doing a ***pilot*** (a small practice experiment) to see if people actually make errors first.

Controllability
Although controlling the complexity of the text and questions between conditions is simpler than that of pictures, you have to be careful. Length and complexity of text and sentence have to be matched, as well as the familiarity of the words/concepts introduced in both. Other factors to take into account are plausibility of the text and the plausibility of the question, given the text. Additionally the question itself may help participants to answer correctly. If this is true for one condition but not another, differences in accuracy do not necessarily reflect differences in understanding the preceding sentence. For example, take the sentence/question pairs in (2).

(2) a. The cake was eaten by the child.
Did the cake eat the child?
b. The dog was bitten by the cat.
Did the dog bite the cat?

Clearly it is much more obvious that the cake did **not** eat the child than that the dog did not bite the cat. Even in the less obvious version in (3a), there is still a problem.

(3) a. The cake was eaten by the child.
Did the child eat the cake?
b. The dog was bitten by the cat.
Did the cat bite the dog?

If participants remember that the sentence contained *cake*, *child* and *eat*, they can answer the question in (3a) correctly. This is not true of (3b).

A drawback of asking comprehension questions is that this task also has a memory component: the participant has to remember the text while processing the succeeding question. This may make this task rather difficult to use with some participant populations (aphasics, for example). Picture verification is less hampered by memory considerations, since the picture and sentence are in different input modalities and can be processed simultaneously, or the sentence can even be presented after the participant has had some time to examine the picture.

2.1.3 Grammaticality Decision

Grammaticality decision requires participants to decide whether sentences are grammatical or not.

Specificity

When people fail to understand the structure of a sentence, they may decide that it is ungrammatical. Thus grammaticality decision can be used as a measure of the ease of understanding a certain type of sentence.

Sensitivity
As with the other tasks mentioned so far, this task is only sensitive to rather large differences in comprehensibility. Furthermore, participants tend to try to understand sentences instead of rejecting them as ungrammatical. That means that you need to include filler items which are clearly ungrammatical in the task to encourage participants to response 'ungrammatical'. The speed of making the grammaticality decision is actually more sensitive than the accuracy, since participants tend to realize that the sentences are grammatical eventually. As with the other accuracy tasks, a pretest will give an impression of how sensitive the task is and how many participants and sentences you will need to detect a difference.

Controllability
The factors that are most important to control in grammaticality decision are those that disturb the ease of making the decision. If a sentence is implausible, participants may decide that it was incorrect and equate that with ungrammaticality. It is therefore important to match sentences in plausibility. If participant do not recognize a word, they may also indicate that the sentence is incorrect, so very low frequency words are to be avoided.

2.1.4 Shadowing

A task that has been used very successfully particularly by Marslen-Wilson (1985) is speech shadowing. In this task sentences are presented over earphones and participants are asked to repeat them aloud, as quickly as they can.

Specificity
The type of errors that people make, the conditions under which they make errors can be provide insight in how people identify words. Marslen-Wilson (1985) used this task to show that almost all errors fit into the semantic and syntactic context in which they are uttered; this suggests that syntactic context has a very quick influence in word identification.

Sensitivity
Although this task can be used without experimental manipulation, as in the Marslen-Wilson (1985) study, the results are very difficult to interpret. However, Marslen-Wilson and Welsh (1978) used it in a more controlled manner by introducing errors of various sorts into the sentences and measured how frequently participants were able to fluently restore the original sound. They were thus able to show effects of various sorts of context on word recognition.

Controllability
What has to be controlled for in a shadowing task? First, the mispronunciations cannot be different from the good pronunciations in delivery speed or emphasis. It is also important that the same sounds be used in the various conditions, otherwise any effects could be due to phonetic perceptibility or produceability of the two conditions. Additionally, if you wish to

test the effects of context, you need to carefully match the contexts in all aspects except for what is being manipulated – relevant factors will be length of words, position in the syntactic structure, familiarity of the words, plausibility, etc.

Obviously this task is limited to being used with certain questions, but is an interesting possibility for those questions.

2.2 Rating Tasks

In a rating task, participants rate sentences on a scale. This scale can be binary ('good'/'bad') or n-ary (a scale from 1 to 7, 1 to 12, or even larger). The most common form of rating is plausibility rating. One other form that has been used is correctability (how easy would it be to correct an ungrammatical sentence to a grammatical sentence with equivalent meaning).

Rating tasks are most commonly used to ***pretest*** experimental materials (Chapter 3). A pretest provides data on which you can match the materials you are going to use in your actual experiment. In some situations, a rating can also be used as an experimental task in itself.

2.2.1 Plausibility

In a plausibility rating, participants are asked to decide whether the situation described in a particular sentence is plausible or not.

Specificity
Plausibility ratings are most frequently used in pretesting of materials to be used later with another task to make sure that the conditions are matched on plausibility. A rating test will provide the experimenter with plausibility scores for each sentence(version). This allows the experimenter to selected and combine experimental sentences such that the conditions do **NOT** differ in plausibility.

Sometimes a rating task may be suitable for an experiment in itself. The prediction then concerns the plausibility of different structures/contexts. An example is an experiment carried out by J. Brouw (1995), using materials like the following (translated from Dutch):

(4) a. The viewer won the challenge.
b. That was the challenge that the viewer won.
c. The challenge annoyed the viewer.
d. That was the viewer that the challenge annoyed.

As you can probably sense, the sentence in (2d), where the object of the relative clause verb is animate and the subject is inanimate, seems very implausible. This clearly is not due to the plausibility of the proposition itself, since (2c) is plausible. The difference in plausibility between (2a) and (b) is much less marked, so the effect is not simply due to the difference in complexity of the two structures.

Another application of plausibility ratings is to show that two groups of participants show a large difference in plausibility rating for a particular structure, but not for others, presumably because one group cannot comprehend that structure correctly.

Sensitivity
Plausibility ratings are not very sensitive. This task is particularly prone to differences between participants. Some participants may find almost everything plausible, while others find nearly everything implausible. In addition, people tend to rate a sentences relative t the last few sentences they read, not relative to all sentences that are in the questionnaire, or sentences that they have encountered in general. Another problem sometimes participants do not remember which end of the scale is plausible or implausible ("1" = plausible or implausible). Good instructions are therefore vital. Participants need to be introduced to the scale very clearly, with explanations of the value associated with points along the scale. Putting the values ('good', 'average', 'bad') along the scale at each item may lead to more consistent responses, as well.

Controllability
Obviously, factors such as structural complexity, or use of low frequency words can lead to a sentence having a lower rank in a plausibility rating. Additionally, if a sentence cannot be fully interpreted without any preceding context, it will seem implausible. For instance, using pronouns (*He, she, that)* can be rather akward if there is nothing that these pronouns refer to.

2.2.2 Sentence Completion

In a sentence completion task, participants are given a sentence fragment and asked to finish it. We classify this task as a rating task, because it provides a way of rating the probability of a sentence ending given the context provided.

Specificity
Sentence completion experiments can be used to investigate the preferred reading of an ambiguous sentence. For example, a completion task can be used with the following fragments to see whether people prefer the relative clause to modify the first NP ('High Attachment'), or the second NP ('low attachment'):

(4) These are the keys to the cabinet that....

Depending on the verb agreement or semantic information in the completion, one can determine which structure the participant assumed for the sentence fragment.

Sensitivity
A completion task allows participant ample time to reflect on the sentence and to change their minds before writing anything down. Psycholinguists however usually would like to know what people's initial interpretation is. Tasks like these are therefore not very specific to investigate what psycholinguists actually want to know.

However, completion tasks can make useful pretests. For instance, say an experimenter wants to investigate whether people use transitivity preference information in determining sentence structure. Say that the experimental items are like those in (6). In both sentences, the boundary between the embedded and the main clause is right after the first verb. The experimenter's intention is to use verbs that are preferably transitive (take a direct object) in one version (6a), and verbs that are preferably intransitive (take no direct object) in the other version (6b).

(6) a. Even though Mr. McKenzie heard his wife didn't appear to notice.
b. Even though Mr. McKenzie cheated his wife didn't appear to notice.

The experimenter therefore needs to be sure that the verbs s/he uses have the intended transitivity preferences. To do that, the experimenter can conduct a completion test in which s/he asks a number of participants to finish fragments like

(7) a. They heard…
b. They cheated…

If a very large proportion of the participants complete a fragment with a direct object, then presumably the structure with an object position is made more accessible when the verb is recognized and is this verb's preferred subcategorization frame. If a fragment is mostly completed without a direct object, then this verb can be assumed to be preferably intransitive. To obtain a good estimate of preferences such as these, one would need many participants do the task (40-60).

Controllability
In a sentence completion experiment, the same factors should be controlled for as with the other rating tasks mentioned above.

2.3 Response Time Measures

In general the problem with the measures discussed up to this point is that they do not permit us to test very detailed predictions. In general, they are useful for testing predictions that one condition will be more difficult than another only when there are large differences. The **time** that participants need to make a response is usually more sensitive than their accuracy: when the sentence is more difficult to understand, it takes longer to make the response, whether we are talking about sentence verification, question answering or grammaticality decision.

Specific predictions about the point at which a difficulty will show up cannot be tested when you are measuring the reaction to a sentence as a whole. Consider Hypothesis 1'', for example. The hypothesis predicts that S-coordinations will be more difficult to understand than subject relative clauses. Moreover it predicts that the difficulty arises when becomes clear that the initially chosen structure was wrong. This is at the verb in the second clause, which is incompatible with an NP-coordination. To really test this specific prediction, you need to look at processing ***on-line***, that is, while the participant is processing a sentence from the first word to the last. The optimal method is therefore one that is a direct measure of the

processing speed. This is difficult to do, since we can't actually get inside participants heads (but see physiological measures below). However, some tasks reflect the processing more or less directly (e.g. eyetracking), whereas with others, the connection between the prediction and the measurement is more indirect (e.g. phoneme monitoring)

Response Time (RT) measures are the most commonly used (because least expensive) techniques in sentence processing experiments with healthy adults. They can be extremely useful also with children and aphasics, although they may have to be carefully fitted to the specifics of the participant group.

2.3.1 Self-Paced Reading and Self-Paced Grammaticality Decision

Self-paced reading is a task where the participant presses a button to get a display (text fragment) on a computer screen. When s/he is finished reading the display, the button is pressed again, providing a measure of the time needed to read the display and get ready to go on to the next. As this technique is the one we will be concerned with most in this course, we will discuss it in more detail than we did the previous techniques.

Specificity and Sensitivity
There are several parameters to self-paced tasks which determine the specificity and sensitivity of the task:

- The size of the text segments to be displayed. Participants can see a word, a phrase, a whole sentence, or even a paragraph at once.
- The nature of the response can be simply reading (Press the button to go to the next segment!), a grammaticality decision (press button x if the sentence still is grammatical, button y if not) or plausibility decision (press button x if the sentence still make sense, button y if not).

Size of the segments
Let us first discuss the question of the size of the text fragments. How do you decide which size chunk to use? First, this is generally a question of what you want to know. If your prediction is that two sentence types differ in difficulty of processing, then presenting entire sentences will yield the kind of data you need. If your prediction concerns the point at which sentences will differ, then presenting a sentence word-y-word or phrase-by-phrase is more appropriate.

A second factor that you might take into account in determining the size of your chunks is to what extent the conditions can be matched in length, frequency, plausibility etc. In general, the larger the chunks to be presented, the more difficult and time consuming it will be to match the conditions adequately. So, having smaller chunks is often simply more convenient than larger ones. In some situations, it may not even be possible to match two conditions if large chunks are used. The contrast you wish to test forces you to take a small chunk size, or otherwise the conditions cannot be compared.

Let us illustrate this latter point with a somewhat elaborated example. Say we want to test the hypothesis that readers pursue only one reading when they read a sentence that is ambiguous. This hypothesis predicts that people take longer to understand a preferred reading compared to a nonpreferred reading. Say you want to investigate this using ambiguous verbs such as 'pushed'. The sentence fragment is (8) is ambiguous in English.

(8) The man pushed into the street
  a. toward the tram.
  b. fell under the tram.

People tend to interpret the fragment as compatible with the ending in (8a), where *pushed into the street* turns out to be a main clause verb phrase. However, at *fell* in (8b) it turns out that it 'push into the street' is relative clause with 'pushed' being the passive participle. We could test the difference in processing of (8a) and (8b) recording the time and accuracy in a grammaticality decision task. Our prediction is that sentences such as (8b) will be responded to more slowly than sentences such as (8a). However, the materials in (8) are **not** matched: in (8a), there is only one clause and in (8b) there are two clauses. There is thus a considerable difference in syntactic complexity. This (and other factors) might affect the difference in response times between the two conditions.

One of the things we need to do is match the sentences in complexity, which means that both conditions need to have two clauses. We can revise our prediction a bit to take care of this problem: a temporarily ambiguous sentence which turns out to have the non-preferred structure will be more difficult to understand than an unambiguous sentence with the non-preferred structure. Now we can use two different conditions, as in (9)

(9) a. The man pushed into the street fell under the tram.
  b. The man who was pushed into the street fell under the tram.

Only unfortunately this will not work if we are recording the reaction times to the entire sentence: (9b) contains more words than (9a). Because reading more words takes more time than reading fewer words, any difference between (9a) and (9b) can be ascribed to a factor other than the experimental manipulation. This is not so bad here, since the prediction is that (9a), the shorter sentence, is more difficult than (9b), the longer sentence. However, that means that the predicted difference is less likely to show up! Apparently, ambiguous and unambiguous relative clauses cannot be completely matched in English, rendering the comparison unsuitable for a task in which reaction times are recorded on sentences presented as a whole. However, if we used word by word, or phrase by phrase presentation, we can measure only the disambiguating region *fell under the tram* without needing to worry much about the extra reading time for the relative clause itself.

A third issue in determining the size of your chunks is reducing variability in your data. The larger the chunk, the more time people spend reading it. If the display consists of a few words, some people will look back and forth between these words. These participants will have a longer RT than people that do not. This means that the variance is larger when using a

large chunk size than when using a smaller chunk size where people cannot look back. Since your aim is to reduce noise as much as possible (Chapter 1, Section 7), you may prefer to use the smallest chunk that is reasonable given your hypotheses.

A final factor that may figure into your decision of the chunk size is *naturalness*. Reading in larger chunks is more natural, since that is how text is normally displayed.

Nature of the response
Now let us turn to the nature of the response the participants have to give. Which kind of task is the most appropriate for your purposes? In general, grammaticality decision is the most sensitive in detecting processing difficulty. This is because participants are forced to pay attention to and make a decision at every word. As soon as they encounter a sense of difficulty, they will need to stop and think whether the sentence is merely ungrammatical, or whether it can still be continued in a grammatical way. This means that in a word-by-word version, effects are likely to be seen at the critical word itself. Simple reading is less sensitive, since it does not force participants to fully process the segment before going to the next word. The result is that processing difficulties may only show up after the critical word. Effects that show up at the next word or later positions are called ***spill over effects***.

In deciding which task to use, the question of naturalness can play a role as well. Grammaticality decision is anything but natural. It slows people down and may make processing more difficult, since the participant needs to carry out judgment task in addition to reading. The effect that you register with this technique may therefore be somewhat artificially induced and may not occur under more natural circumstances.

In sum, when you choose self-paced tasks, you will need to take into consideration your specific hypothesis and prediction, the variability generated by larger segment sizes, and the naturalness of the task, and be able to argue why you decide to use one task over the other.

2.3.2 Eye Tracking

Eye trackers register the eye's position on a screen or other type of display, and can yield a very extensive record of what participants look at and how long for. Eye tracking has been used to study text reading, and to study people's eye movements in response to spoken text. We will focus on the first line of research here.

Specificity
Eye tracking is a very direct measure of the time necessary to read a particular word or phrase in a particular context, which makes it a very good measure in sentence comprehension. Eye tracking also provides a rather natural reading situation, in which people can view a large section of text and can freely move their eyes back and forth. On the other hand, it is sometimes difficult to decide how to define your reading time measure in this sort of data. Some generally used measures are: how long the reader initially looks at a word (first pass reading times); after having fixated the word, how much time the reader takes to re-read a preceding region before reading the word again (regression times); how long the

word is read when the reader comes back to it after having read something else in the sentence (second pass reading times), and the total time spend reading a particular word (total reading times).

Sensitivity
Eye movements are sensitive to a number of properties of the words (length, frequency, fit into the context, etc), and eye tracking been used very successfully in sentence processing research. However, there are a number of potential pitfalls.

Controllability
Eye movements are very sensitive to visual factors that do not depend on the sentence level of processing at all. Such factors are of lesser importance in other techniques, simply because other techniques are less sensitive to visual processing. This means that if you are planning an eye-movement study, conditions must be strictly matched on several parameters (word length in characters, length of preceding word, cf. Trueswell, Tanenhaus and Garnsey, 1994). In addition, short words, such as determiners, are usually not fixated, since they can be seen from the preceding or following fixation. One should therefore avoid using very short words as the critical words in an experiment. Another issue to bear in mind is line breaks, where people move their eyes from the end of one line to the beginning of another. Reading a phrase than spans a line break will take much longer than reading a phrase that does not. Critical regions should therefore be matched on this parameter, or better: not include any line breaks at all.

2.4 Indirect and Concurrent Tasks

Sometimes it is difficult to use a task that directly tests the hypothesis, but a concurrent measure can provide an indirect measure. One such indirect measure tests ***interference***: the more difficulty that participants have with the main task, e.g. sentence comprehension, the more difficulty they will have carrying out the concurrent task.

2.4.1 Lexical Decision/Naming

In lexical decision experiments, people are asked to decide whether a letter string is a word or not. In naming experiments, they are asked to say a word out loud. The speed and accuracy of the decision or naming can be measured.

Specificity
Lexical decision and naming tasks are generally used to investigate the factors affecting word recognition. However, they can be applied to sentence processing in three ways.

The first is to determine the effect of the sentence context on the lexical decision /naming. An example of this is an experiment done by Marslen-Wilson and Tyler (1980). They asked participants to say a word that they saw on a computer screen. This was after ambiguous sentence fragments as in *(10).*

(10) a. Visiting relatives ARE
b. Visiting relatives IS

The word in upper case letters had to be named. The word fits into one meaning (relatives who are visiting) or the other (going to visit relatives). They found that the naming speed reflected the structure which participants had chosen for the preceding fragment. The rationale is that the fit of the word within the preceding context affects the naming times of that word. The task can thus be used to check which syntactic structure is assigned to preceding sentence fragment.

Tanenhaus et al. (1979) and Swinney (1979) used a lexical decision task to investigate sentence processing. They reasoned that if a word is presented in the context of a semantically related word, people recognize it more quickly (semantic priming) and can therefore name it or make a lexical decision more quickly, than when it is preceded by an unrelated word. Many experiments have shown this effect when words are presented in lists or pairs. Tanenhaus et al. wanted to determine whether all meanings of an ambiguous word were accessed during processing or only the meaning that fits into the preceding sentence context in sentences like those in *(11)*.

(11)	a.	The people all rose.	FLOWER
	b.	She smelled the rose.	FLOWER
	c.	The people all rose.	SHOWER
	d.	She smelled the rose.	SHOWER

The context should be able to decide between the verb and noun meanings of *rose*. If only the verb meaning ('to rise') is available in (10a), people should react slower to *FLOWER* than in (10b) where the entire context is semantically related to *FLOWER*. But if all meanings are accessed, then words related to both meanings will show semantic priming relative to the unrelated words in (10c and 10d). They found priming after both (10a and 10b) immediately after the ambiguous word is heard, suggesting that people maintain both readings of an ambiguous word.

Lexical decision or naming can also be used as a concurrent task (a task that takes place at the same time as the primary task). The reasoning is that the more difficult the primary task is, the more it will interfere with the concurrent task, in this case lexical decision or naming. If a sentence is difficult to process, making the lexical decision while processing the sentence will also be slowed down, since the resources necessary to carry out the lexical decision are being used by sentence processing. This task in its truly concurrent form was used by Clifton, Frazier and Connine (1984) to investigate the effects of transitivity expectation. They found that lexical decision was delayed when the sentence did not match the usual transitivity pattern of the verb.

Controllability

Which factors need be controlled for depends on which kind of indirect task you are using. In a priming task such as Tanenhaus et al.'s, you must be very careful that the related words (*FLOWER* in the example) are exactly comparable to the unrelated word (*SHOWER*), except

for the relatedness with the preceding context. This means matching the words on the number of syllables, number of characters, frequency or even visual similarity. The context sentences also need to be matched.

2.4.2 Phoneme Monitoring

Another task which has been used as a concurrent task is phoneme monitoring. Here participants are asked to press a button as quickly as possible when they hear a certain sound ('monitor for a phoneme'), e.g. [p] (Foss, 1969). The rationale is similar to the other indirect tasks discuss above: if sentence processing gets harder, people would need to detect a particular sound and respond to it. As with other experimental techniques, there are many factors that need to be controlled for, e.g. the position of the sound in the word, the predictability of the word containing the sound (Cutler, 1981).

2.5 Physiological Measures

There are several physiological measures that are interesting, because they potentially give us a direct look at what is going on in the brain. They have some drawbacks. In this section we will briefly discuss ERPs, PET and fMRI.

2.5.1 Event Related Potentials (ERPs)

Brain cells (neurons) communicate by transmitting small electric currents. When a large group of neurons are active at the same time, the change in electrical potential can be detected by placing electrodes on the scalp. Psychologists and psycholinguists are interested in how and when the electric potentials change in reaction to the presentation of a certain type of stimulus. In psycholinguistic research, this stimulus usually is a particular word or speech sound in a particular type of context.

The signal generated in response to one trial (e.g. one word) is rather small compared to the other on-going brain activity, which is not related to the stimulus. However, if a large number of trials are averaged, the effects that not related to the stimulus will reduce, since their direction is random. The stimulus-related (and hence, non-random) response will become apparent through averaging. ERPs are therefore obtained by recording brain activity (EEG) while stimuli are presented, and averaging the chunks of EEG starting from the presentation of each stimulus. ERPs can be easily disturbed by eye movement, muscle movement or other factors. In practice, you would therefore need a large number of stimuli to obtain a good ERP signal (advisably, at least 30 items per participant per condition).

Specificity and Sensitivity
The ERP is an averaged wave form which reflects small positive changes and negative changes over time of the electrical potential. Usually, experimenters will present their participants with two or more types of stimulus, and the ERP generated by different

conditions is compared. There are three ways in which wave forms can differ, which allows us to say whether processing is different for the conditions or not. The first is the time point at which the wave forms start to differ. Second, the direction of the difference (negative or positive) is informative. Third, the distribution of the difference over the scalp is important. If two conditions yield waves with a different scalp distribution, one can conclude that the neurons generating the wave form have a different location and/or orientation in the two conditions (although the localization of these neural sources cannot be determined with ERPs).

Much of the language research done with ERPs has focussed on the difference between semantic and syntactic processing. Kutas and Hillyard (1980) showed that semantically unexpected words, such as *socks* in (12b), produce a negative deflection of the wave form when compared to expected words such as *butter* in (12a).

(12) a. She spread the warm bread with butter.
b. She spread the warm bread with socks.

This negative deflection is largest about 400 milliseconds after presentation of the incongruent word. This component is therefore called the N400 component. As for the distribution across the scalp, this negativity is most prominent a little to the right of the midline of the head and is largest at central to posterior electrodes. One way to interpret this difference is the amount of effort needed to access word meaning and to fit it into the semantic context. The semantic information about a highly predictable word is mostly already implicitly available from semantic interpretation of the context.

Syntactic problems, on the other hand, tend to lead to an earlier negativity and a later positivity. Neville et al (1991) compared sentences like those in (13).

(13) a. The scientist criticized John's proof *of* the theorem.
b. The scientist criticized John's *of* proof the theorem.

When a word appears in a position where it does not belong syntactically, as in (13b), there is a negativity at about 125 msec after the word appears on the screen. This negativity to be bigger over the left hemisphere than the right, hence the name Left Anterior Negativitym, or LAN. In addition, words that introduce a syntactic difficulty elicit a positivity at around 600 msec after presentation, called the P600 component. It is very widespread over the head, but has a posterior maximum. This component is very commonly found with both ungrammatical sentences and sentences in which people have initially picked the wrong structure. It has been described as indexing difficulty in integrating the word into a structure.

We can argue that the LAN, N400 and P600 components represent different sorts of processing, (i) because they differ in time course (125 msec, 400 msec, 600 msec or later), (ii) because they differ in polarity, and (iii) the scalp distribution differs, suggesting again that the underlying neuronal generators are different (although we cannot determine their locations very accurately on the basis of scalp distribution).

Controllability
ERPs are sensitive to the plausibility of a word (cf. N400), its length and other visual components. It is therefore best to use the same words in the contrasting conditions. In addition, it is sensitive to repetition, so it is better to use a content word only once per participant. Furthermore, the linear position of the word in the sentence also has been shown to affect the ERPs. As we have already mentioned, a large number of materials is necessary to obtain relatively 'clean' ERPs. Constructing materials is therefore a fairly difficult job.

2.5.2 Brain imaging: Functional Magneto Resonance Imaging (FMRI) and Positron Emission Tomography

2.5.3 Brain Imaging Techniques

Positron Emission Tomography (PET) and functional Magneto Resonance Imaging (fMRI) are method used to measure and localize activity in the brain. In psychological and psycholinguistic research, blood flow is primarily used to investigate the brain regions involved in particular cognitive activities: brain areas that are more active while performing a certain task will require more blood. Hence cognitive processing can be investigated by looking at the differences in regional cerebral blood flow between two or more conditions.

PET Using PET, blood flow can be detected from the outside of the body by injecting a small amount of radioactively labelled water into a participant's vein and placing the participant with his/her head in a PET camera. On the basis of the pattern in which the gamma rays are emitted and detected by the camera, the location and amount of blood flow in the brain can be computed.

fMRI fMRI is less invasive, and does not require injection of any material. Instead, it registers the reaction of the body's molecules to rapid changes in a strong magnetic field. Blood containing oxygen has different magnetic resonance properties than blood without oxygen. If an areas of the brain becomes more active in a certain task, it will get more blood with oxygen and less without oxygen. Changes in the magnetic resonance signal can therefore tell us which parts of the brain are active in one experimental condition versus another.

Aside from being less invasive than PET, fMRI has another advantage, namely that one can take pictures of the brain every 2 seconds or more often. In PET it often takes longer than 1 minute to obtain a picture. This is not to say that fMRI can track rapid cognitive activity: it takes about 6 second after the onset of a task or stimuli before the bloodflow pattern has changed. Brain imaging techniques thus have a very limited temporal resolution compared to the techniques discussed above.

Specificity
PET and fMRI are appropriate techniques to use when you want to know what areas of the brain are involved in a particular cognitive activity. For instance, some researchers use brain imaging to argue for certain distinctions claimed for in cognitive models, for instance the distinction between semantic and syntactic processing. If research shows that syntax and

semantics are carried out in different areas if the brain, this theoretic distinction is supported. Another application of brain imaging methods is to track reorganization of the brain after lesions or during development. Both brain imaging techniques are extremely sensitive to body motion, which makes it hard to use this technique in children or patients. Another drawback is that using these techniques is rather expensive. Imaging centers may charge $400-800 per hour of scanning.

Controllability
As is the case with all experiments, brain imaging requires a tight matching of the conditions. Even though fMRI and PET may be less sensitive to small differences between stimuli, as an experimenter you want to make sure that any differences you find in bloodflow between the experimental conditions cannot be attributed to other factors.

2.6 Concluding Remarks

In this chapter, we have discussed some of the major techniques used in psycholinguistic research, ranging from off-line accuracy measures to physiological measures such as ERP and brain imaging. There is no optimal technique. Which technique is best depends in the first place on your hypothesis. Other considerations are how easy it is to make materials given the sensitivity of the technique, and o f course, whether the technique is available your institution and whether you can afford using it from a financial point of view.

Chapter 3 Constructing the Right Materials: Avoiding Artifacts and Reducing Noise

Making up materials is a confounded nuisance -Anne Cutler

As has been made clear in Chapter 1, what you're trying to obtain as the result of your experiment is a significant **difference** between two or more experimental conditions. The materials which you use have to be constructed with two things in mind. The first is that the only difference between the conditions should be the experimental manipulation per se: in AB's experiment, this is NP versus S coordination. To ensure that this is the case, when constructing your materials you need to make sure that you have controlled for all other differences which might otherwise cause **artefactual differences** between the conditions. Second, you need to be aware of the variability within your materials which may cause **noise**. One of the factors that determine whether a difference is significant or not is the variability of the results as expressed by the variance or standard deviation: if each participant responds to the experimental condition in a different way or if a participant responds to each sentence in a condition differently, the variance of the responses within the conditions is large. This means that it will be more difficult to pick up a significant difference relative to another condition. On the other hand, if the variance is small, even a small difference between two conditions could become significant.

Recall the *archaeopteryx* example from Chapter 1, Section 6. The experimenter's hypothesis was that NP coordinations (Condition A) will take less time to read than S-coordinations (Condition B). One of the test sentences in condition A contained the unfamiliar word *archaeopteryx*, leading to a large variability of the reaction times for Condition A. So, even if the experimenter's hypothesis was correct, the difference between the two conditions might have been obscured by the large variance. The use of an unfamiliar word thus introduced some ***noise*** in the data, which prevented results from being significantly different.

Now suppose that the sentences in Condition B included some unfamiliar words, instead. Suppose that the experimenter found that Condition B was reacted to significantly more slowly than Condition A. Does that mean that her hypothesis was correct and S-coordinations (B) are harder to process than NP-coordinations (A)? No! Although a difference between A and B was predicted, the difference found between A and B could be attributed to other things than the type of coordination, namely the familiarity or frequency of the words in Condition B. The result obtained could thus be an ***artifact***, that is, an effect due to a flaw in the experiment, rather than a reflecting a real difference.

Noise prevents an effect from being significant; artifacts cause a significant effect, but for the wrong reason. Artifacts and noise thus render your results hard to interpret: on the basis of the experimental results nothing can be concluded concerning the hypothesis you started out with. So, in designing and running an experiment, the things you really want to control are factors that may cause artifacts and noise. In this, and the following chapters you will learn what these factors are and how you can reduce their potential effects on your data. In Section 3.1 we

will deal with the construction of experimental sentences. The order of presentation will be discussed in Section 3.2. Section 3.3 deals with factors other than materials or presentation.

3.1 Constructing Materials in Sets

Once you have chosen an experimental technique and know exactly which conditions you are going to compare, your next step is constructing the items your participants are going to read or listen to. Ideally, the items in your experimental conditions should be exactly the same except for the things you are intentionally making different in order to carry out your test. As mentioned in the introduction of this chapter, any other difference might cause artefactual results.

For purpose of illustration, let us consider AB's first experiment again, which was dealt with in Chapter 1. According to AB's Hypothesis, temporarily ambiguous sentences that turn out to have an non-preferred structure (S coordination in this case) will be more difficult to process than sentences that turn out to contain a preferred reading (NP coordination). AB wants to test this by recording the response time in an end-of-sentence grammaticality judgment task in which the sentences are presented as a whole. AB therefore decides to look at several types of temporarily ambiguous coordination. Her prediction is that sentences in which 'and' coordinates two NPs (preferred structure, Condition A) will be responded to faster than sentences in which 'and' coordinates two clauses.

Suppose that AB's sentences look like the ones below. Clearly, this is NOT the way to test the hypothesis. Why not?

Sentences in Condition A (NP coordination):

I saw the man and the girls that came out of the hotel last night.
The little old lady asked the pilot and the passenger for directions.
Kim kissed Tom and me.
Down came the vase and the books.

Sentences in Condition B (S coordination):

John ate the burger and I devoured the hotdog.
The landlord mowed the lawn and his son swept the porch.
Laura picked up Peter and we went to Antarctica together.

Clearly a difference in reading times between sentences in condition A and B could be due to dozens of other factors besides the preference for one type of coordination over the other:

- the length of the item varies between and within the conditions
- differences in syntactic complexity are also found within the conditions and between conditions
- the conditions differ drastically in the information that allows the reader to determine which reading is intended by the speaker
- there seem to be noticeable differences in semantic plausibility of the sentences.

Some of these differences are systematic between conditions and may thus cause artefactual differences: which? Some are equally obvious within conditions and thus cause noise in the responses: which? Note that number of sentences is not even the same across conditions!

The easiest way to control for most artefactual differences between conditions is to construct ***sets of materials***. An ***item set*** is a group of items which are exactly the same except for the experimental manipulation. In addition, all sets will be rather similar in syntactic complexity, number of words, and other issues that may lead to an increased variability in reaction time. In AB's experiment, some sets may be the following. Note that the comparison between the A and B conditions is still not ideal.

(1)
Set 1:
Condition A: The hairdresser recognized the grocer and the doctor in the store.
Condition B: The hairdresser recognized the grocer and the doctor spotted the plumber.

Set 2:
Condition A: The dentist greeted the nurse and the patient in the office.
Condition B: The dentist greeted the nurse and the patient greeted the receptionist.

Set 3:
Condition A: The painter sanded the chair and the table late last night.
Condition B: The painter sanded the chair and the table is already finished.

With sentence sets such as (2) above, can all eventual differences between the conditions be attributed to a preference for NP-coordination an S-coordination? Clearly not. We already touched upon most of them in Chapter 1. Below we will discuss the following factors in somewhat more detail and show how you can control for them:

- (word or sentence) length
- frequency of word occurrence
- plausibility of the sentences
- repetition of words and sentences

This list is not exhaustive. Other factors can also affect the response times of the participant (what about imageability and animacy?). The experimenter has the responsibility for always considering what variables may yet be unmatched between her or his conditions. Since variability within your materials can never be completely eliminated, any factor that can be matched between conditions should be matched. In this way, this factor will merely cause noise (and hopefully as little of that as possible) and will not lead to an artefactual result.

3.1.1 Matching Conditions on Word Length

It is a well-known fact that long words, such as *antidisestablishmentarianism* take longer to read than short words such as *an*. Length clearly has an effect on response times. So, if one condition includes only long words, and the other condition relatively short words, you are likely to find an artefactual difference between the two conditions. Under these circumstances, you do not know which factor to attribute a difference to: your intended experimental difference or length.

What do we mean by ***length***? That depends on how the materials are presented, as well as on the task to be carried out. If the material is presented visually, and participants only have to read it, length can be defined as the ***number of characters***. If the material is presented auditorily or if participants have to name words, the ***number of syllables*** is a more appropriate definition.

How to control for length? The ideal way would be to make pairs of words, one for each sentence set that are of exactly the same length. However, in many experiments this is impossible. For instance, in AB's paradigm (2) above, in Condition B the critical disambiguating word will always be a verb (*spotted, greeted, is*), and hence, will often be longer than the word in the corresponding positions in condition A (*late, in*).

For most experiments, it will suffice to find pairs of words such that the mean length (and standard deviation of the length) of the words used in one condition is similar to the mean length (and standard deviation) of the words used in the contrasting condition. If this is the case, the two conditions are matched in length, even though the critical words for each particular item set may not be. You can additionally test to what extent the conditions are matched by applying simple statistics, such as a t-test or an ANOVA. Note that the difference should be as insignificant as possible, hence the p-value should be as large as possible (e.g. $p>.2$).

The procedure discussed in this section provides a match between conditions to avoid artifacts. Additionally, choosing to limit yourself to fairly short words throughout the experiment is useful in order to eliminate noise from the data set.

3.1.2 Matching Conditions on Word Frequency

As you know from the *archaeopteryx* example in Chapter 1, the frequency of a word also has an effect on a participant's responses. A word that is very frequently used, such as, for instance, *the* or *bread* in English is recognized and processed faster than somewhat less common English words such as *jumper* or *bereft*. So, the crucial words in the experimental conditions not only have to be matched on length, but also on their ***frequency (of occurrence)***.

Fortunately, electronic databases are currently available and accessible in which frequency information can be found. One such database is the Brown corpus, originally created by Francis and Kucera. This data base can be electronically accessed through the MRC database (e.g., http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm). Databases such as these give several kinds of information for each word: its orthography (spelling), word class (e.g., noun or verb), pronunciation, length, and its frequency of occurrence, among other things.

In the Brown corpus, the frequency data are based on a sample of texts that contains 1 million word tokens. These texts are drawn from all kinds of written materials (newspapers, novels, magazines…) to make sure that the vocabulary in the database is representative of the normal range of language use. For each word form listed, the Brown Corpus provides the number of times this particular word form occurred in this collection of 1 million words (which is called the **raw** frequency). Some frequency databases are based on larger corpora, for instance, the English database in Celex (http://www.ru.nl/celex/) is based on a corpus of 17.9 million words. The British National Corpus (http://www.natcorp.ox.ac.uk/) is based on a corpus of 100 million words. The raw frequency provided by these data bases is always the number of occurrences in this particular data base (e.g. a raw frequency of 30 in Celex, means '30 per 17.9 million words'). To be able to compare frequencies across corpora, it may be useful to calculate the frequency per 1 million. Some databases already provide this frequency as an option.

In addition to the raw frequency, psycholinguists often report the ***logarithmic frequency***. Logarithmic (log) frequency generally gives a better estimate of the psychological impact of frequency differences between words. Take for instance, a noun such as *beaker*. According to the Celex data base, this has a frequency of 33 in 17.9 million words. Now, consider *basket*. This word is probably much more familiar to you. The frequency of this word is 428 in 17.9 million, 395 units more than the frequency of *beaker*. Now consider some high frequency words. Let's take *woman*. This word form has a frequency of 15241. Now consider *child*. Can you guess if it is more or less frequent than *woman*? Hard to tell by your intuitions. However, according to the frequency counts, *child* has a frequency of 19389, that is, 4148 more than *woman*! Clearly, your intuitions about frequency do not correspond to the linear scale of the raw frequency of a word. A few units make intuitively a large difference in the case of low frequency words, but for high frequency words, several hundreds of units do not seem to matter so much. That is why logarithmic frequency is often used in matching stimuli. By calculating the logarithmic frequency, the linear scale of the raw data is transformed into an exponential one. As a result, small differences in the low-frequency range will have a large impact, whereas large differences in the high-frequency range will have a small impact. Let us illustrate this with the examples given above. To calculate the log frequency, the log base 10 is calculated of the frequency per 1 million words. The word *beaker* has a frequency of 33/17.9= 1.8 per 1 million. The log frequency is log base 10 of 1.8, equals 0.27; *Basket* has a log frequency of 1.38, that is, a difference of more than 1 unit. *Woman* has a Log frequency of 2.93, and *child* 3.03, a difference of 0.1 unit only! The frequency as represented on a logarithmic scale thus corresponds more to your intuitions than the 'raw' frequency represented on a linear scale. In developing and matching your experimental materials, it is therefore advisable to use the logarithmic frequency of the words.

Another important distinction is that between ***word form frequency*** and ***lemma frequency***. Form frequency is the number of occurrence of a particular (inflected) word form. For instance, the verb form *says* has a word form frequency that is different from the word form frequency of *saying* or *said*. Lemma frequency, on the other hand, is the frequency of all forms of a particular word taken together. For instance, the lemma frequency of the verb *to say* is the sum of the frequency of *say, says, said* and *saying*.

Which kind of frequency, form or lemma, do you need for matching your materials? That depends on the kinds of words that are crucial to your experiment. In general, lemma frequency will be more reliable than word form frequency, as the former is based on more data. If the various different forms of the word do not differ much in meaning and frequency (for instance, singular and plural noun forms), you could use the lemma frequency to match your materials. However, if the frequency of the forms of a word differ a great deal, as in the case of the different verb tenses, word form frequency is more appropriate. For instance, the form *dies* is far less frequent than *died* or *dying*. Hence, if you are comparing present tense verbs in your experiment, you should consider word form frequency. Which kind of frequency you choose to match your material therefore depends on the differences between the experimental conditions and the questions you are interested in. The most important thing is to think over your options clearly and to know why you decided on lemma or word form frequency.

Note that the Brown Corpus frequency counts are based exclusively on written, and hence, somewhat formal materials. As a result, words that sound very common such as *puppet* are relatively low in frequency compared to abstract terms such as *justice*. For this reason, familiarity ratings of words (accessible through the MRC database (http://www.psy.uwa.edu.au/mrcdatabase/uwa_mrc.htm or collected by the experimenter on a large enough sample of the targeted population) can be more accurate than mere frequency data. However, for normal adult speakers, the Brown corpus data will suffice, especially if you restrict your materials to words with a frequency of more than 1 per million. However, things are different if you are testing non-native speakers or children. For these participant groups different frequency corpora and/or familiarity rating data should be used.

The frequency data provided by electronic data bases allows you to match your conditions on frequency to avoid artifacts. Ideally, the mean and standard deviations of the frequency should not be significantly different between the conditions. Frequency information can also be used to minimize noise in the experiment, by rejecting extremely high or low frequency words.

3.1.3 Matching Conditions on Plausibility

Not only length and frequency, but also factors that cannot be determined on the basis of a database have to be matched between the conditions. Suppose that you are interested in the difference in processing of nouns and verbs at the sentence end. Suppose that your stimulus material includes the following sentence set:

(4)
Condition A: John ate the bananas and apples.
Condition B: John ate the bananas and sneezed.

Clearly it is somewhat uncommon that people sneeze after eating a banana, while eating several fruits at the same time is not uncommon. Because of this, your participants will react differently to the second sentence than to the first. Hence, the reaction times for the two conditions may well differ, and not only because there is a noun in the first and a verb in the second!

So, conditions also have to be matched on plausibility. But how do you measure plausibility? Plausibility is not like the numbers of occurrence in a corpus, or the number of characters, and there are no databases from which you can easily derive this information. Instead, you have to build your own thermometer to measure the plausibility of a sentence. This thermometer is a group of people that rate your (potential) experimental sentences on a plausibility-scale. This is most easily done in a paper-and-pencil rating task (see Chapter 2, section 2), using a non-binary scale. The plausibility of a test sentence is the mean score this particular sentence received in the rating.

Such a rating or ***pretest*** is actually a mini-experiment on its own. This means that when you are conducting a pretest, you have to be concerned about issues that are important for every experiment, such as repetition, fillers, latin square and other things to be discussed below.

It is wise to include more sentences in the pretest than you plan on using in the actual experiment. This way, you can pick out the very best sentences (i.e. sentences with the highest mean plausibility rating and smallest variance) to use in the actual experiment and discard the worst. This will minimize noise from differences in plausibility. It is also important not to use the people you have asked to rate your sentences for plausibility in your actual experiment. They will certainly remember some of the sentences! As will be discussed in the next section, repetition effects have to be avoided.

3.1.4 Effects of Repetition

A researcher could try to avoid effects of length, frequency or plausibility by presenting a perfectly matched pair of sentences over and over again. But imagine what a participant would do if he or she had to read a list of sentences like this:

(5) The hairdresser recognized the grocer and the doctor in the store.
The hairdresser recognized the grocer and the doctor spotted the plumber.
The hairdresser recognized the grocer and the doctor in the store.
The hairdresser recognized the grocer and the doctor spotted the plumber.
The hairdresser recognized the grocer and the doctor in the store.
The hairdresser recognized the grocer and the doctor spotted the plumber.
The hairdresser recognized the grocer and the doctor in the store.
The hairdresser recognized the grocer and the doctor spotted the plumber.

Etc.

Not only will the participant get bored, he or she will also begin to read faster and become less attentive to the experimental manipulations. Furthermore, the participant may start to guess the next word which is coming up based on the earlier versions of the sentence, and start wondering why the experimenter is interested in 'and' sentences. All this leads to an abnormal way of sentence processing that might obscure the differences in normal processing which you wish to find. So, repetition has to be avoided. This included literal repetition of the sentence, repetition of structure and repetition of lexical items.

When participants read (nearly) the same sentence over and over again, they will memorize it. This sort of repetition should be avoided at all costs. When participants begin to consciously predict the next word(s) based on memory, it will affect normal sentence processing.

Second, it is a problem to repeat the same syntactic structure too frequently. Participants will begin to pay too much attention to the structure rather than reading the sentence normally. For this reason it is important to disguise the experimental structure as much as possible.

Third, not only should entire sentences not be repeated, also repetition of content words (lexical verbs, nouns and adjectives) in a different context has to be avoided, as this might introduce effects that disturb your experimental comparisons. It is known that participants remember words from an experimental list that they have seen some days (even months) earlier. Repetition leads to changes in brain potentials (ERPs) and in regional cerebral blood flow (fMRI and PET). So in constructing your sentence sets, you should avoid repeating words, since this will introduce noise. At the very least, you should match the number of repetitions across your experimental conditions. To experience the effect of lexical repletion, let's consider a version of the example sentence sets:

(2)
Set 1:
Condition A: The hairdresser recognized the grocer and the doctor in the store.
Condition B: The hairdresser recognized the grocer and the doctor spotted the plumber.

Set 2:
Condition A: The dentist greeted the hairdresser and the doctor in the office.
Condition B: The dentist greeted the hairdresser and the doctor greeted the receptionist.

Set 3:
Condition A: The painter saw the dentist and the grocer late last night.
Condition B: The painter saw the dentist and the grocer saw the plumber.

Hairdresser, dentist, grocer, doctor and *plumber* are used in two sentence sets; this will cause confusion and draws attention to the experimental manipulation. We would be better off by using different nouns.

You may wonder what good has it done to make sentence sets then if they cannot be presented to a single participant because of repetition effects? We will return to this point in the Section 3.2.1, when we discuss participant lists.

3.1.5 Reporting on Materials Development: Expressing the Match between Conditions

Once you have collected information concerning length, frequency, plausibility, and other factors, you can start matching your experimental conditions. In the ideal situation, your conditions should differ on none of the factors considered above. For your and other researcher's convenience, you should be able to provide a table containing all means, standard deviations and

the outcomes of statistical tests to show that your conditions do not differ (the difference should be non-significant (n.s.)). Such a table might look as follows:

	Length		Log.Freq		Plaus	
	Mean	SD	Mean	SD	Mean	SD
Condition A:	6.6	3.4	1.47	.07	6.8	1.2
Condition B:	6.5	3.1	1.45	.10	6.7	1.0
t-value:	t(6)= 1.0		t(6)= .99		t(6)= 1.0	
significance:	n.s.		n.s.		n.s.	

A convenient way to match your materials is to enter your data in Excel or another spreadsheet program. This will make it easy for you to calculate means and standard deviations, shift words and sentences around between conditions, and calculate the means again until the best match is found.

3.2 Presenting the Materials; or How to be Smarter than Your Participants

We have talked about the factors to be taken into account while constructing materials in order to avoid artifacts and to minimize noise. The purpose of this section is to discuss the actual presentation of the materials, and the problems that can be avoided, if due care is taken during the construction of the test lists. The first step is to avoid repetition effects by assigning versions of sentence sets to different lists. The second step is disguising the purpose of your experiment by using distractors items, or ***fillers***. The third step involves the choice of the order in which the materials are presented.

3.2.1 Avoiding Repetition Effects with an Optimal Match: Distribution of Sentence Sets over Lists

As discussed in the previous section, experimental materials need to be constructed in sets in order to minimize noise and artifacts. This is easiest to do by making sets of sentences which are identical except for the experimental manipulation. However this apparently is at odds with the ban on repetition discussed in section 3.1.: If a single participant sees every version of the sentence set, there will be a rather large number of repetitions. Let us illustrate this with AB's second experiment, and show the way out of this dilemma.

Say that AB decided on an experiment using four conditions: Sentences containing a temporarily ambiguous S-coordination in which the NP after 'and' is inanimate (Condition A), a similar type of sentence in which the NP is animate (Condition B), and sentences containing an S-coordination, disambiguated by a comma, in which the NP after 'and' is inanimate (Condition C) or animate (Condition D). Say that AB has constructed four sentence sets:

(6)

Set 1:
a. The hairdresser recognized the scissors and the comb was also familiar.
b. The hairdresser recognized the scissors and the doctor was also familiar.
c. The hairdresser recognized the scissors, and the comb was also familiar.

d. The hairdresser recognized the scissors, and the doctor was also familiar.

Set 2: a. The painter finished the walls and the ceiling was almost done.
b. The painter finished the walls and the plumber was almost done.
c. The painter finished the walls, and the ceiling was almost done.
d. The painter finished the walls, and the plumber was almost done.

Set 3: a. The lady noticed the stains and the rug was rather smelly.
b. The lady noticed the stains and the dog was rather smelly.
c. The lady noticed the stains, and the rug was rather smelly.
d. The lady noticed the stains, and the dog was rather smelly.

Set 4: a. The janitor grabbed the mop and the bucket was fetched immediately.
b. The janitor grabbed the mop and the principal was fetched immediately.
c. The janitor grabbed the mop, and the bucket was fetched immediately.
d. The janitor grabbed the mop, and the principal was fetched immediately.

Let's suppose that these materials are all:

- Matched for the frequency of the noun after 'and';
- Matched for the length of the noun after 'and';
- Pretested, selected and matched for plausibility of the situation expressed.

Of course, AB could present all the sentences above to each of her participants. Each participant would then receive the following list:

(7) The hairdresser recognized the scissors and the comb was also familiar.
The hairdresser recognized the scissors and the doctor was also familiar.
The hairdresser recognized the scissors, and the comb was also familiar.
The hairdresser recognized the scissors, and the doctor was also familiar.
. The painter finished the walls and the ceiling was almost done.
The painter finished the walls and the plumber was almost done.
The painter finished the walls, and the ceiling was almost done.
The painter finished the walls, and the plumber was almost done.
The lady noticed the stains and the rug was rather smelly.
The lady noticed the stains and the dog was rather smelly.
The lady noticed the stains, and the rug was rather smelly.
The lady noticed the stains, and the dog was rather smelly.
The janitor grabbed the mop and the bucket was fetched immediately.

Etc.

Clearly, this is not the right way to present the sentences. As mentioned in section 3.1.3: participants will get bored, become less sensitive to the difference and start developing all kinds

of guessing strategies. Clearly, this is not what she wants to happen, so, she'd better avoid participants reading more than one member of each sentence set. However, all four members of one set still must be tested: only by doing this can differences in lexical items (frequency, length, plausibility) be adequately controlled for. Additionally, using item sets provides a better way to test generalizability to items (a within-item statistics, cf. Chapter 5).

The way out of this dilemma is to construct different ***participant lists*** of materials; each list contains only one version of a given sentence set, but enough lists are made such that each version appears on one list. Each participant sees only one list, so that s/he only sees one version of each experimental sentence, but an equal number of items per condition. The general procedure for doing so is called a ***Latin square design***.

3.2.1.1 Participant lists and Latin square design

The investigator must take care that none of her participants sees more than one item from the same set of sentences. There are four conditions in AB's experiment, hence four items per set, so four different participant lists, I, II, III and IV, must be constructed. Suppose that 28 participants take part in the actual experiment. In this case, seven participants will read list I, seven list II, seven list III, and the remaining seven list IV. In this way, each version of the item set contributes to the experimental result, but no participant sees more than one version of an item set and AB gets an equally good estimate of the response for each version of the sentence by having equal numbers of the participants read each list. Now, what is the best way to distribute the sentences across the lists? Let us start with set 1. AB can put 1a in list I, 1b in list II, 1c in list III and 1d in list IV. What about the next sentence sets? As there are four sentence sets, each participant must see 4 (sets) / 4 (conditions) equals 1 sentence in each condition. One way to adequately distribute the sentences across the participant lists is the following:

	Participant List			
	I	II	III	IV
Sentence version	1a	1b	1c	1d
	2b	2c	2d	2a
	3c	3d	3a	3b
	4d	4a	4b	4c

In this way, each list contains one sentence in each of the conditions, and no list contains more than one version of each sentence. This way of distributing items across participant lists is called a ***Latin Square design***.

If you have successfully constructed your sentence sets with appropriate matching of factors such as length and frequency, with no repetition between participant sets, then producing a set of lists of this format means that there should be **no artifacts** in the experimental design, including repetition artifacts.

3.2.1.2 Matching between item groups: Minimizing noise continued

In practice, you may want to have each participant respond to more than one sentence in each condition, say 8 for a reading study, or 40 for a study using event-related brain potentials. Including more than one item per condition will give you a better idea of what people are doing. In addition, you want to be able to generalize the results across items. In general, it is best to have a Latin Square with an equal number of items in each condition on each list. Hence, if you have four conditions, you should construct a multiple of four item sets.

Suppose AB had constructed 16 sentence sets. In this way, four sentences could appear in each condition on each list as shown below. We'll get back to what "group" indicates.

	Participant list			
	I	II	III	IV
group (i)	1a	1b	1c	1d
	2a	2b	2c	2d
	3a	3b	3c	3d
	4a	4b	4c	4d
group (ii)	5b	5c	5d	5a
	6b	6c	6d	6a
	7b	7c	7d	7a
	8b	8c	8d	8a
group (iii)	9c	9d	9a	9b
	10c	10d	10a	10b
	11c	11d	11a	11b
	12c	12d	12a	12b
group (iv)	13d	13a	13b	13c
	14d	14a	14b	14c
	15d	15a	15b	15c
	16d	16a	16b	16c

Recall that **over all participants** the effects of length, repetition and plausibility etc. are controlled, since the conditions are matched within the sentence sets. However, the individual participants only see one of the lists, that is, only one version per item set. Hence, if you randomly assign versions to a list, like we did in the above, it is not guaranteed that the sentences each participant sees (that is, *within* each participant list) are matched between conditions. This may introduce noise in your results. Say that item sets 13, 14 and 15 happen to be less plausible than other sets. Participants reading list I will therefore have an artificially increased response time for condition D; participants in list II for condition A, etc. Recall that in

order to see a significant difference between conditions, all participants in the experiment must react in more or less the same way to the experimental conditions. The misbalance in plausibility may therefore increase the variance of the reaction times, which may obscure potential differences between the conditions that are of interest.

The solution is to first make item groups, match these in terms of factors such as length, frequency of the critical items and plausibility, and use these groups to make the participant lists. Notice that sentence sets 1-4; 5-8; 9-12 and 13-16 above belong together: In each list, sentences from sets 1, 2, 3 and 4 are seen in one and the same condition (a in I; b in II; c in III, and d in IV). Similarly, sentences from each of the other groups always appear in the same condition across the lists. That is to say that the sixteen sentence sets are divided into four ***item groups***: group (i) containing sets 1-4; group (ii) containing sets 5-8; group (iii) containing sets 9-12; and item group (iv) containing sentence sets 13-16. If these groups are not matched, the result will be noise in the experimental results.

These groups should be matched on length, frequency, plausibility and other factors that might affect a participant's response (see section 3.1). In this way, not all difficult items are concentrated in one item group, and, hence, in one condition per list, but are equally distributed across the conditions instead. As a consequence, length and frequency effects will affect all conditions within and between lists in a similar way. As we will see in Chapter 5, statistical analyses are done on the mean response time for each condition, so that if the participants' mean RT for each condition is affected equally by all these sources of variance then the only factor left to cause a difference is the experimental manipulation and a significant difference between conditions is more likely to show up.

In practical terms, you can check the match between item groups by making a table showing how the item groups (and thus the conditions on each list) are matched on a variable such as lengthy, frequency etc. The table below illustrates plausibility scores. Notice that the four groups are matched within each list (columns) and that the lists (rows) are very much like each other as well. This means that each participant should show a very similar effect of plausibility and the noise due to this factor is as small as possible.

	List I	List II	List III	List IV
Group(i)	a=3.9	b=3.8	c=3.8	d=3.9
Group (ii)	b=3.8	c=3.9	d=3.75	a=3.9
Group (iii)	c=3.85	d=3.75	a=3.9	b=3.85
Group (iv)	d=3.9	a=3.85	b=3.85	c=3.7

The procedure for matching and the way to do this using is the same as the procedure used for matching the experimental conditions, as described in Section 3.1. The differences between the groups and lists can be statistically tested by running t-tests or, with more than two conditions, One-way ANOVAs on each row and each column. Again, you are looking for a combination of item sets within item groups that will yield a difference that is as insignificant as possible, hence you want extremely high p-values.

3.2.2 Fillers as Distractors

Latin squaring solved the problem of repetition of the items with each participant. However, presenting only experimental items is not ideal. Even without repetitions, participants will be likely to begin noticing the pattern in the sentences, leading to conscious analysis and possibly to unusual strategies. This list can be made more exciting to read by including some non-experimental items, ***distractor items,*** or ***fillers***, that differ from the experimental items in structure. The advantage of inserting fillers is not only that they make participants more attentive, but also that they (i) disguise the purpose of the experiment, (ii) can serve as a check on whether a participant is correctly doing the task, and/or (iii) can serve as a separate ***pilot*** experiment.

A rule-of-thumb is that a list contains twice as many fillers as experimental items. So, if your experiment consists of 4 conditions with 10 sentences each per participant list, a reasonable number of fillers is 40 x 2 = 80, yielding a total of 80 + 40 = 120 sentences. If a participant sees 120 sentences of various types, it is harder to guess which 40 sentences you are interested in, and if your participant is clever enough to single out your experimental manipulation, he or she will only figure it out towards the end of the experiment.

Why is it so important to have the participant be naïve concerning the purpose of the experiment? If participants know what the experiment is about, they start parsing the sentences in a different way: they start making predictions about the materials to come, or they do not process the semantics of the sentence but only look at the sentence format, or develop other kinds of reading strategies. However, what a psycholinguist tries to investigate is the human language processing mechanism in its most natural state, and not the cleverness of the people participating in the experiment. For this reason, one should also **never tell a participant what *exactly* your experiment is about until after the experiment**.

A second function of fillers is that these items can serve as a check on the way participants are completing the task. Suppose you are conducting an off-line choice task in which participants have to choose one of two possible continuations of a sentence. For instance, participants have to indicate whether they prefer a singular or a plural noun in sentence fragments like

(7) He claimed that books / book …

Whether or not a participant is consistently and seriously completing the task can be checked by including some filler items in which only one choice is possible, such as:

(8) I saw ten girl / girls…

The number of errors on these items will indicate how carefully your participants were doing the task.

A third potential reason for including fillers is efficiency. If two experiments are not too similar, the two experiments can be conducted at the same time, with the materials of one experiment serving as filler materials for the other. Sometimes, fillers are used for a pilot experiment: a small 'test'-experiment, often with not perfectly balanced materials, to see whether there is any

indication of a predicted difference between conditions. On the basis of the outcome of such a pilot, the investigator can decide whether it is worth the time and effort to design and run an experiment with a full set of perfectly balanced materials, and obtain an estimate of how many more materials are likely to be necessary.

In constructing filler materials one should take care that the syntactic structures of these items used do not resemble the actual experimental materials too much: otherwise the task will still be rather boring, participants will still easily guess what you are investigating, or will find a certain processing strategy, and your data will be hard to interpret. On the other hand, if the experimental materials are completely different from the fillers, they will stick out like a sore thumb. If, for instance, your experimental sentences are 12 words in length, using filler stimuli of only 5 words in length will not help distracting attention from the experimental materials.

Here is an example of the problems of using a too similar structure. Suppose the goal of your investigation is to tap the so-called reduced relative garden path effect. Your experimental sentences are of the type (9a) and (9b) below:

(9) a. The horse raced past the barn fell.
b. The horse that was raced past the barn fell.

In the a-versions, verbs like *raced* are used that are temporarily ambiguous between a passive participle and a simple past tense. The b-sentences are their unambiguous counterparts in which the structure is immediately obvious. Since reduced relative clauses as in (9a) are relatively uncommon, we expect readers to initially understand the a-sentences as main clauses. In such an experiment, you would do better not to include unambiguous reduced relative clauses such as (9c) in your filler materials.

(9) c. The letter written to Mary disappeared.

In this case, the verb *written* can only be a passive participle, and the reduced relative interpretation is the only option. Reading many sentences of type (9c) is likely to remind participants of the existence of the structure and thus make it easier to choose it also in the ambiguous cases than it normally would be. Because of this bias, the experimental sentences of the type in (9a) will become easier to parse. The garden path effect in sentences such as (9a) will be reduced, and hence the difference you are interested in, namely the difference between (9a) and the unambiguous relative in (9b) will become smaller.

Repetition of structures **within** the fillers is advisable. Constructing fillers of a limited number of different structures will make it somewhat easier to come up with filler sentences. Furthermore, it will make it even harder for the participant to figure out what the experiment is about: instead of seeing one sort of structure that shows up repeatedly, they will see several different types of structure repeated. Repeated filler structures thus serve as distractors. For example, if you need 80 fillers to distract people from S-coordinations, you might construct 20 fillers of structure *w*, e.g. sentences with a conjunct different from 'and'; 20 of structure *x*, e.g. sentences starting with an embedded clause; 20 of structure *y*, e.g. sentences containing a embedded clause; and 20 of

structure z, e.g. sentences with an 'or' NP conjunct (see next section for examples). There is a good chance that the participants will not notice that you are interested in S-coordinations.

For reasons illustrated below, it is easiest to have the number of filler sentences for each filler type be a multitude of the number of sentences per experimental condition. For instance, if you have 3 experimental conditions, and each participant will see 12 items per condition, it is easiest to make 12, 24 or 36 sentences of each type of filler.

In 3.1 we talked about avoiding repetition of words in your target sentences. For the same reasons, the filler sentences should not contain words which are also used in the experimental sentences: if you use the same words in your fillers and experimental materials, the responses of the participants in the experimental conditions will be prone to repetition effects.

3.2.3 Pseudorandomization and Scrambling

A third issue in presenting your experimental materials concerns the order in which the sentences are presented. Let's return to AB's materials; for simplicity's sake we will take the set of four sentences. Suppose AB followed our advice and included (4 x 2 =) eight filler sentences, two each of four types, in her material, yielding the following participant list I.

(10) The hairdresser recognized the scissors and the comb was also familiar.
The painter finished the walls and the plumber was almost done.
The lady noticed the stains, and the rug was rather smelly.
The janitor grabbed the mop, and the principal was fetched immediately.
The actress bowed, but the audience did not give her a standing ovation.
The cat hissed, but the bulldog did not leave the yard.
While the librarian was sorting out the books, the girl walked over to the boy.
Because the cubs were tired and hungry, the lioness laid down in the grass.
The realtor mentioned that the owner and the renter were not on good terms.
The teacher said that students and parents could not enter after 8pm.
The flight attendant or the pilot addressed the passengers before take-off.
The chicken or the rooster has been making a lot of noise lately.

Clearly, a list such as this is not what we wish to present to our participants: all sentences of a similar type are presented next to each other! This is the most likely way to make participants notice what the structures are. To make a participant list somewhat more attractive to read and to distract attention from the target items, fillers and experimental sentences must be ***scrambled***, or ***randomized***.

Furthermore, recall that 'real' experiments often use more than one sentence per condition, which must be spread over the entire list to avoid learning and effects of fatigue. We also wish to avoid having more than two sentences of a similar kind next to each other. Therefore, we can't simply close our eyes and arbitrarily pick the first, second, third, etc. sentence. Instead, the order of presentation must be ***pseudorandom***, that is, apparently arbitrary, but not really so.

What we want is a list in which items from each condition are spread out across the list with fillers of each type interspersed through them. How can we accomplish this goal?

Pseudorandomizing by creating blocks

One way to pseudorandomize the order of the experimental materials is to assign the sentences to ***blocks*** in the following way. Recall that the experimental materials are divided in to several item groups, that is, those sentence sets that are read in one and the same condition in a certain list. For AB's experiment with 16 item sets, the item groups were the following (for convenience we only show the conditions of list I):

group (i)	1a
	2a
	3a
	4a
group (ii)	5b
	6b
	7b
	8b
group (iii)	9c
	10c
	11c
	12c
group (iv)	13d
	14d
	15d
	16d

Furthermore, AB has four different kinds of fillers, as described above. These are represented as filler group i, filler group ii, filler group iii and filler group iv below. Since she has 32 filler sentences (2 x number of experimental items (16) equals 32) there are actually eight sentences in each filler group, but we only give the first two of each group.

Fillers (i):[1] 1. The actress bowed, but the audience did not give her a standing ovation.
2. The cat hissed, but the bulldog did not leave the yard
:

Fillers (ii): 1. While the librarian was sorting out the books, the girl walked over to the boy.
2. Because the cubs were tired and hungry, the lioness laid down in the grass.
:

Fillers (iii): 1. The realtor mentioned that the owner and the renter were not on good terms.
2. The teacher said that students and parents could not enter after 8pm.

[1] You may want to use different item numbers to keep the fillers distinct from the experimental items, e.g. 101-108 for the first filler group, 201-208 for the second, etc.

:

Fillers (iv): 1. The flight attendant or the pilot addressed the passengers before take-off.
2. The chicken or the rooster has been making a lot of noise lately.
:

To ensure that the experimental conditions and filler types are not clustered together in the participant lists, but are evenly distributed across the list, we construct a series of blocks of materials which occur together within the list. To create a block, choose one sentence from the first item group (i.e., from one condition, for example Condition A on list I), then one from the second item group (another condition), one from the third item group (a third condition), then one from the fourth group (a fourth condition), and you have the experimental sentences belonging to one block. Make blocks until all the experimental sentences have been allocated. That means, in this case, that you have made 4 blocks of four sentences each. In this way, the conditions are spaced out over the list, not clustered together.

You also wish to distribute the various types of fillers across the participant list. Typically, there will be more sentences of each type of filler than there are blocks. It is convenient to have the number of sentences per filler type be a multitude of the number of blocks (i.e. number of experimental items per condition in each participant list), so that each block can contain an equal number of fillers of each type. Simply assign one sentence from each different filler group to each of the blocks you have just created. For instance, the first sentence of filler group (i) will be assigned to the first block, the second sentence of filler group (i) to the second block, etc. Repeat until all the sentences of filler group (i) have been assigned to a block; then turn to filler group (ii). Bring some variation in the assignment, e.g. assign the first sentence of filler group 3 to block 3, the second to block 1, etc. In this way the fillers will be spread evenly, but somewhat randomly across the blocks. Given the number of filler sentences (8) of each filler type (4) in AB's materials, and the fact that she has four blocks, each block will contain two of each filler type. The first block (on list I) might contain:

The hairdresser recognized the scissors and the comb was also familiar < item group i
The painter finished the walls and the plumber was almost done. < item group ii
The lady noticed the stains, and the rug was rather smelly. < item group iii
The janitor grabbed the mop, and the principal was fetched immediately. < item group iv
The actress bowed, but the audience did not give her a standing ovation. < filler type i
While the librarian was sorting out the books, the girl walked over to the boy. < filler type ii
The realtor mentioned that the owner and the renter were not on good terms. < filler type iii
The flight attendant or the pilot addressed the passengers before take-off. < filler type iv
The cat hissed, but the bulldog did not leave the yard. < filler type i
Because the cubs were tired and hungry, the lioness laid down in the grass. < filler type ii
The teacher said that students and parents could not enter after 8pm. < filler type iii
The chicken or the rooster has been making a lot of noise lately. < filler type iv

The whole participant list I (containing four blocks) would look something like this. Fillers are denoted by the number of their group followed by their own number within the group (see next page).

Block 1	Block 2	Block 3	Block 4
1a	3a	2a	4a
6b	5b	7b	8b
11c	10c	12c	9c
14d	16d	15d	13d
i-1	i-2	i-3	i-4
i-6	i-7	i-8	i-5
ii-2	ii-3	ii-4	ii-5
ii-3	ii-4	ii-5	ii-6
iii-7	iii-8	iii-1	iii-2
iii-5	iii-6	iii-7	iii-8
iv-4	iv-5	iv-6	iv-7
iv-8	iv-1	iv-2	iv-3

Notice that we did not simply assign the first sentence in an item group or filler type to block 1; the members of the item and filler groups are randomly spread across the blocks. This is to prevent possible effects due to the way that you have chosen the sentences to put in the item groups (shortest to longest, for example) and the order of making the sentences (from easiest to make up to hardest, for example, which probably corresponds to hardest to understand as well!).

Randomizing the order within blocks

We have still got the problem we noticed earlier: experimental items, though not of the same condition, still immediately follow each other. (Still it is a lot less extensive than if the 16 sentences and 32 fillers were presented in an order where each type of sentence is presented together!) Furthermore, the order of the four experimental conditions is always the same across the four blocks; the fillers are not doing their work of distracting the participant if they all come at the end of the block. This order will certainly arouse suspicion in clever participants. What has to be done is to scramble the order of fillers and experimental items ***within each block***. We can simply throw all the sentences (within a given block) up in the air and pick them up in whatever order they come to hand, and that is the order in which they will be presented. AB's list I will now look like this:

Block 1	Block 2	Block 3	Block 4
14d	iii-6	iv-2	8b
iii-7	i-2	15d	i-4
iv-8	5b	ii-5	iii-2
ii-3	16d	i-3	ii-6
iv-4	i-5	12c	13d
6b	iv-5	2a	iii-8
i-6	10c	iv-6	9c
11c	3a	ii-4	iv-3
1a	ii-3	i-8	ii-5
i-1	iv-1	iii-7	4a
ii-2	ii-4	iii-1	iv-7
iii-5	iii-8	7b	i-1

Block 1 would look like:

The janitor grabbed the mop, and the principal was fetched immediately.
The realtor mentioned that the owner and the renter were not on good terms.
The chicken or the rooster has been making a lot of noise lately.
While the librarian was sorting out the books, the girl walked over to the boy.
The flight attendant or the pilot addressed the passengers before take-off.
The painter finished the walls and the plumber was almost done.
The cat hissed, but the bulldog did not leave the yard.
The lady noticed the stains, and the rug was rather smelly.
The hairdresser recognized the scissors and the comb was also familiar.
The actress bowed, but the audience did not give her a standing ovation.
Because the cubs were tired and hungry, the lioness laid down in the grass.
The teacher said that students and parents could not enter after 8pm.

Now even the cleverest participant will have a hard time figuring out what you're testing!

Randomizing the order between participants

There is still one potential problem left. Suppose that the order above is the only order in which you present your items. Suppose you have very clever participants who can figure out what you're testing, or –a more likely situation- that your participants get tired during the test. Or, -something which might happen in a grammaticality decision experiment - that your participants learn over the list not to decide too quickly that sentences are ungrammatical… In all these cases, sentences at the end of the list will be processed in a different way than the sentences presented at the beginning of the list, even if they are structurally the same.

Recall that you also wish to generalize across items (Chapter 1, Section 1.6). That means that you want people to react in the same way to all the sentences in a given condition. If participants reacted as you expected only to the first sentences on the list, and not the last, due to the factors just mentioned, you are unlikely to get a consistent enough pattern to reach signifi-

cance: the items presented at the end of the list will be different from those presented at the beginning and the item analysis will not reach significance due to this noise.

To balance the effects of fatigue, learning, boredom or cleverness across all your experimental sentences, also the order **between blocks** must be scrambled, such that sentences sometimes appear in the first, sometimes in the middle and sometimes in the last part of the experimental list. If you have *n* blocks, the number of different orders you can create per participant list is *n*!. That is, if you have four blocks, the number of possible orders is 4!= 4 x 3 x 2 x 1 = 24. When using stimulus presentation programs, such as Eprime, you can have the program scramble the order between and within blocks in a different way for each participant. In off-line experiments, where one has to scramble by hand, making a different order for each participant is hardly realizable, and one has to be satisfied with a few orders of each participant list.

3.3 Other Ways to 'Beat' Your Participant

Fillers help a lot to distract your participant's attention from the purpose of the experiment and keep your participant engaged in processing the sentences in a more or less natural way. An additional method to prevent boredom is to have your participants perform an additional task. For instance, a participant can be requested to answer comprehension questions about sentences during or at the end of the task. Alternatively, participants can be asked to recall words, sentence fragments or react to a certain target word or letter while reading/listening to sentences (as in a phoneme monitoring task). One must however bear in mind that an additional task affects the way participants process the materials. For instance, if participants are asked to remember the final words of each sentence, it is likely that participants pay less attention to the other words of the sentence, or the content of the sentence as a whole, than if they are required to answer comprehension questions.

3.4 Summary

In constructing the experimental materials, you should control for factors that may cause artifacts and noise. If you neglect these factors, your results will be hard to interpret.

To avoid **artifacts**, you must match your materials between conditions on syntactic structure, length, frequency, plausibility and other factors that might influence a participant's response. The most efficient manner to match between conditions is to construct a series of **item sets**. An item set is a **set of items which do not differ except for the experimental manipulation**. Length and frequency data can be obtained by consulting the electronic databases such as the Brown corpus or CELEX. For less objective factors such as plausibility, a rating must be carried out. Repetition is another factor that might influence your data, and hence, has to be avoided. Where a close match between conditions is not possible, sentence sets should be carefully balanced so that the conditions remain matched over all.

To prevent **repetition effects**, each version of an item set must be assigned to a different **participant list**, such that each list **contains only one member of an item sentence set**, and the

same number of sentences in each condition. The general procedure for doing so is called a **Latin square design.**

Additional measures need to be taken to reduce the effects of **noise** (increased variability in responses). Creating participant lists means that each participant does not see the total set of materials, and, hence, that the conditions are not matched per participant list (although they are matched in the experiment as a whole). This will not cause an artifactual result, but can cause noise that prevents the experiment from producing a significant result. To minimize potential effects of noise, it is important to choose **item groups**, which are **groups of item sets that appear in the same condition within every list**. If the item groups are matched as closely as possible on factors such as length, frequency and plausibility, the mean response of the participant to each condition will only reflect the experimental manipulation.

Ad-hoc processing strategies, habituation, boredom, and/or fatigue on the part of the participant may also produce noise. These effects may be reduced by interspersing **filler items** or **distractors** among the experimental sentences (to distract from the purpose of the experiment), and by presenting the experimental items and fillers in a blocked pseudorandom order. This means that one item in each condition and a number of fillers of each type are randomly assigned to a **block of sentences which occur together in the list.** The order within the block is randomized. This will prevent items of the same type to be presented together, and will help to further disguise the goal of the experiment. In addition, the participant's attention can be distracted from the experimental manipulation by a secondary task (e.g. answering comprehension questions about the sentences) or special instructions. Never tell your participant exactly what you are testing before the experiment! Notwithstanding all caution, the effects of boredom and special strategies can never be really avoided. To be sure that such effects will equally affect all your stimulus material, the order of blocks and order of items within blocks should be scrambled individually for each participant, when technically feasible.

Chapter 4 Running Your Experiment

In the previous chapters you learned how to construct materials, choose a measurement technique, and control factors that might lead to artefactual results or noise. The next step is to present the participant lists to a group of people and collect their responses (e.g. reading times, responses to comprehension questions, etc.).

Presenting your materials is rather easy when using questionnaires: you simply hand out some booklets containing your experimental sentences (and instructions!), and after the participants have completed the task, you collect the booklets again and look at the responses. On-line experiments such as self-paced reading or self-paced grammaticality judgement tasks, however, cannot be run without the help of a computer and software to control stimulus presentation and data acquisition. The experimental software we will use in this course is the Eprime software. Below, we will give a very brief introduction on how to modify existing scripts.

In the second part of this chapter (Section 4.2), we will deal with the actual experimental procedure. We will show that even in this stage of the experiment, noise has to be avoided by using the right experimental environment and by properly selecting, instructing and training your participants. Last, but not least, we will stress the importance of keeping good records (logging) and making back-ups.

4.1 Importing your items in E-Prime

4.1.1 Coding Items

Before you enter the items into an E-prime script, make sure all items have insightful codes associated with them. These codes will help you keep track of which response belongs to which item, and will make it easier to prepare data for statistical analysis.

For instance, say you are going to conduct an experiment in which participants see one entire sentence at a time, and will need to press the 'Q' key for 'Correct' and 'P' for 'Incorrect.' The items in Participant List 1 may then be coded (in Excel) in the following way:

Itemnr	Cond	Itemgroup	Sentence	Correct
14	d	4	The janitor grabbed the mop, and the principal was fetched immediately.	Q
307	f3		The realtor mentioned that the owner and the renter were not on good terms	Q
408	f4		The chicken or the rooster has making a lot of noise lately.	P
203	f2		While the librarian was sorting out the books, the girl walked over to the boy.	Q
404	f4		The flight attendant or the pilot will addressed the passengers.	P
6	b	2	The painter finished the walls and the plumber was almost done.	Q
106	f1		The cat hissed, but the bulldog did not leaving the yard.	P
11	c	3	The lady noticed the stains, and the rug was rather smelly.	Q
1	a	1	The hairdresser recognized the scissors and the comb was also familiar.	Q
Etc.				

List number 2, for instance, would start like:

Itemnr	Cond	Itemgroup	Sentence	Correct
14	b	4	The janitor grabbed the mop and the principal was fetched immediately.	Q
307	f3		The realtor mentioned that the owner and the renter were not on good terms	Q
408	f4		The chicken or the rooster has making a lot of noise lately.	P
203	f2		While the librarian was sorting out the books, the girl walked over to the boy.	Q

Etc.

The only real difference between the lists is in the conditions of the experimental items. It is not illustrated here, but another difference between the lists may be the order in which the stimuli are presented.

Note that the item numbers of the fillers are greater than 100, to distinguish them from the experimental items (see Footnote 1). It is CRUCIAL that you insert a column with the condition code. In addition to condition codes such as 'a', 'f4' etc as illustrated above, you can insert one or more columns that describe the conditions, e.g. 'Animate', 'comma'. In this way, it is easy to keep track of which condition code stands for what experimental manipulation.

4.1.2 E-Prime

In most on-line behavioral experiments, you want to record how long it takes for a participant to make a response to a particular stimulus, and what kind of response is being made. For instance, in a self-paced experiment using a word-by-word moving window display, you want to know the amount of the time that has elapsed from the moment a word appears on the screen to the moment the participant presses the button. One of the software programs that can be used to run such experiments is E-prime, which we will use in class. Other software packages are, among others, PsyScope, DMASTR (freeware!) and Presentations.

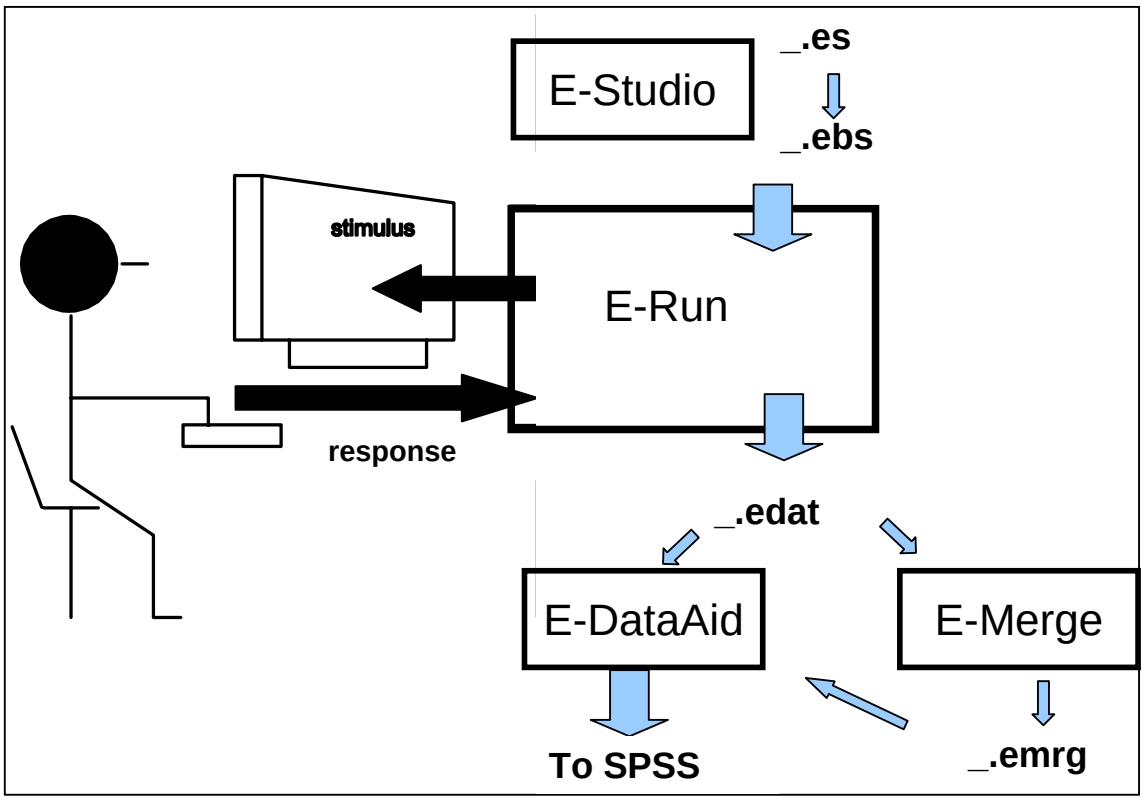
E-prime is a collection of programs (E-Run, E-Studio, E-DataAid and E-merge) that will help you create and run an experiment.

E-Studio. You can specify how you want the stimuli to be presented and what responses you want to record, in E-Studio. You will use E-Studio to insert your list of stimuli into the experiment. The stimuli and instructions are saved in an _.es file. After clicking 'run' in E-Studio, you can see what the experiment looks like to the participant, and automatically create an _.ebs file.

E-Run When running participants, it is best to use E-run. Load the _.ebs file into E-Run and click 'run'. E-Run. An _.edat file will be created, containing the stimuli, the actual timing of the stimuli, and the timing and type of participant responses.

E-Merge. E-Merge can concatenate (string together) two or more _.edat files. This is convenient if you run more than one participant. E-Merge will create an _.emrg file.

E-DataAid. E-DataAid can read both _.edat files and _emrg files. This program allows you to reorder, delete and filter columns and rows in the _.edat/.emrg files, and allows you to export the data to a format that can be read into Excel or SPSS.



We will not discuss the structure of an E-Studio file and how to make a new experiment from scratch. In most cases, you will only need to slightly modify existing scripts. Please consult the E-Prime manual and website for instructions and examples.

4.1.3 Importing stimuli into E-prime _.es files

After you have constructed your participant lists (cf. Chapter 3), the next step is to insert these stimuli into E-prime. The easiest way to do this if you have multiple participant lists, is to create a different _.es file for each list. Stimuli are inserted in List Object in the E-Studio _.es file. It will contain the columns:

ID WEIGHT NESTED PROCEDURE

The first three columns can be ignored for now. 'Procedure' will contain the name of the procedure that specifies how the stimuli will be presented. For instance, you can insert 'TrialProc' in the rows. TrialProc will be defined in the next object and consists, for instance, of a fixation screen, a screen presenting the first word, a screen presenting the second word, etc.

Your stimuli need to be inserted in the List Object.

- First create as many columns as you need to accommodate the Stimuli, the item number, the group number, the condition code, and expected (correct) response, if applicable. Do this by clicking the icon "Insert attributes".
- Next, rename the columns by clicking on the "Attribute" header of each column.
- Then, copy and paste the rows from the Excel file containing your items into the appropriate columns in the List Object

Displaying the items

- In the TrialProc, or other procedure, refer to the columns in the List Object by inserting the column name between '[]'. For instance, say that you want to present an entire sentence at once, and have the sentences listed in the column "Sentence". Trial Proc will then contain a Text Display with the content '[Sentence].'
- To control the presentation time and mode, click on the 'properties' icon on the TextDisplay, → Duration/Input
- To make sure the responses are recorded, click on the 'properties' icon on the TextDisplay, → logging

Always run all lists before you start running 'real' participants, and check and double check that the stimuli are displayed correctly and that all responses you need are being recorded.

4.1.4 Output files: _.edat

E-prime will generate an output file (_.edat) for each participant you run. This _.edat file will have as name the _.es/_.ebs file that was run, plus the participant and session number you entered when you ran the _es/_.ebs file in E-Studio or E-Run. Data files can be opened in E-DataAid. You will see that a great number of columns are listed that contain information about randomization, the timing of the stimuli etc. Most relevant are of course, the columns containing the item number, condition code, item group, response times and response accuracy. E-Dat allows you to reorder and select the columns you need for analysis. We will practice this in the lab.

4.2 Running Your Experiment

Once your Eprime scripts or questionnaires are ready and checked for typos and other errors, you are ready to 'run' participants and collect data (a truly wonderful moment in the life of a researcher)! Also this requires some preparation. First, we need to find participants and a quiet place to run the experiment. Second, we need to tell the participants what to do (instruct and train them). Only then can we have them read the experimental materials and record their

responses. Other things you need to consider are the logging and back-up procedures. We will discuss these things in turn.

4.2.1 Selecting Participants

First you need to find participants. But you just can't go about and randomly pick people off the street. Recall from Chapter 1 that one of the goals of the experiment is to generalize the results of over a group of people larger than the group of participants that were in your experiment. It is part of the research question and hypotheses to which population you wish your results to be generalizable, for example 'literate adult native speakers of a particular language.'

The participants you choose should form a representative sample of this population. Exclude participants who do not belong to the group. They may disturb your results. If you are not interested in second language acquisition, you better restrict yourself to native speakers of the language of your stimulus materials. Non-native speakers, that is, people who did not learn the language before puberty, might process the language somewhat differently from native speakers. Furthermore, they might not know some words or expressions used in the materials, leading to an increase in reading times, or to unpredictable patterns. One such participant might therefore increase the variance in your data, thus reducing the likelihood that a real difference will be significant.

Finally, you must be sure that your participants are able to fulfill the task. For instance, in reading experiments, you want to avoid participants that are dyslexic or cannot read well from a computer screen, e.g., due to a visual impairment.

Once you have specified which group you intend to generalize over, you should try to match over lists or otherwise control for variables which are known to cause variation among participants, unless your research is directed to these differences. For example, age, sex, handedness, IQ and other respects may be suspected to be influential. If you think one of these factors might influence the results, you should take care that participants with different traits are equally spread over the participant lists. For instance, suppose you expect that left-handers to press the 'NO' button (left side of the keyboard) somewhat more quickly than right handers. If eight of your participants are left-handed and you have four participant lists, you assign two left-handers to each of the four lists. In this way, handedness will affect all items and conditions in a similar way.

In summary, think about which people you wish to include in your participant group and how to assign participants to your participant lists.

4.2.2 Experimental Environment

Not only outlying materials and participants can cause noise, also the experimental environment can influence your results. Preferably, the experiment should be run in a sound attenuated booth.

A sudden knock on the door, a telephone ringing, a motorcycle passing by, or even the experimenter coughing may distract the participant from the task, leading to a temporary increase in reaction times, or to a total loss of concentration if an apparently very funny conversation is being carried on next door. Not only sound, also light circumstances are important: it is hard to read letters on a computer screen when this screen is placed in front of a window on a sunny day. It is just as tiring to read full-bright letters in an otherwise pitch dark room. Also, the participant must be comfortably seated. A chair that is too high or too low, or uncomfortable for any other odd reason will invite participants to turn and move -- and, by Murphy's law, they will be likely do this while processing items that are crucial for your experiment!

So, take care that the experimental environment is quiet in auditory, visual and other respects.

4.2.3 The Experimental Procedure

Once we have recruited our participants and have found a quiet place to conduct the experiment, there are a few more things we need to take care of before we can start collecting data.

Informed consent
First, we need to inform the participant what will happen and what the risks are of participating. By law, a participant first need to agree to the procedures by signing an informed consent form before any data can be collected. Make sure to retain a copy and give another copy to the participant. Informed consent forms need to be approved by the local Internal Review Board (IRB). For information on how to obtain IRB approval, what informed consent forms should look like, and the rationale for the informed consent process, see the IRB website at your local institution.

Instructions
Starting the experiment, we first need to ***instruct*** the participant, that is, we need to explain what s/he needs to do or may not do, how to handle the button box, whether speed or accuracy or both are stressed, etc. Do you trust yourself to remember saying all that to each participant? I don't! It's better to write the instructions down and have each participant read them through. In addition, make a checklist of the major points and go over them with the participant before s/he starts. Make sure that the participants really understand what they are expected to do. You would be surprised how often people do silly things (pressing the 'YES' button when they mean 'NO', for instance).

Practice block
One way to make sure people understand what they need to do is to start with a practice session. Stay with the participant while they are doing this, so you can make sure that they are pressing the right button, sit comfortably, that they can see well, etc. A block of practice sentences typically consisting of 7-10 trials only, and should contain all aspects of the task. For example, if in the actual experimental materials comprehension is tested by means of yes/no questions, the practice block will contain some yes/no questions as well. If a participant is very hard to train, you can go through the practice twice. Or even give up and dismiss the participant (make sure to compensate her/him for the time spent in the lab). Only if you are sure the participant really

understands the task and has gotten used to the way of responding can the actual experiment be started.

After the experiment:
Debriefing
When a participant has finished the experiment, it is often a good idea to ask the participant questions to find out what people are actually doing. Example questions are: 'What do you think about the experiment?' 'Did you notice any problems', ' Did you notice any items that had x (were x is your experimental manipulation)?', 'What do you think the experiment is about?', 'Did you have any strategies?' This is called a ***debriefing***. Record comments in the log book (see section 4.2.5 below).

Backing up the data
Last and most important of all, make a ***backup*** of the data immediately. Lightning could strike your computer at any moment, but if you follow this rule you will not have lost your data.

4.2.4 Running a Practice Participant

We suggested in the last session that you debrief your participants. Sometimes that will tell you something important about a fault (or typo) in your experimental materials or in your procedure. It is rather annoying to find this out when the experiment is nearly finished. It is therefore always helpful to run a practice participant. Find an ultra-critical friend or colleague, preferably. Check whether the machinery is running properly. Check whether your instructions were clear. Check how much time it takes participants to learn the task and how long the practice block should be. Record how much time it takes to run the complete procedure (including instructions and practice block). Carry out an extensive debriefing. Try to nip any problems in the bud before you running 'real' participants.

4.2.5 Logging

It is very important to maintain a log for your experiment containing notes about the experiment and each experimental session. In this way, you can keep track of where you are during the experiment. Have I already run the 5th participant on list III or not? Secondly, it will help you find all the details of the experiment even weeks after you have run your last participant. This comes in handy when preparing a research report in which you need to state all the methodological details (e.g. number of females, average age) of the experiment. Third, it may help you screen the data. Say, for instance, that you notice that one participant has very long reaction times on a few items, or performs only 50% correct. You can then go back to your log and see whether there was anything special about that session or participant. Did s/he fall asleep during the study? Did s/he appear to be spaced out? This is a good reason to drop the participant's data from analysis.

In sum, a ***log book*** is a means of keeping track of what you have already done, what you plan to do, and all the information about your participants that you might conceivably want to report on

later (these are the ***parameters*** of your participant group). An example page from a log book is given below. Consider what other information you might want to keep about your participants?

The logbook acts as a record of which participant this is, what list s/he is assigned to, and as a check list of things that need to be done as well, including the name of the eprime file to be used, the name to be given to the result file, whether the results have been backed up (!), whether the data has been merged and other processing stages. The effort you put into logging really has its rewards!

4.3 Summary

NOT On-line experiments such as self-paced reading or grammaticality judgment tasks can only be run with the help of a computer program that controls the way stimuli are presented and responses are collected. It is very important that your stimuli are coded in a way that indicates which item group each item belongs to. This will allow you to recover which RTs belong to which stimuli and will make further data processing a lot easier.

Before you can really start running your experiment, you have to carefully select your participants and assign them to your participant lists such that potential effects of e.g. handedness will be balanced across the conditions and items. The actual experimental procedure includes the following steps: instructing and training your participant, running the actual experimental files, maintaining a quiet experimental environment to reduce potential noise in your data, debriefing your participants and backing-up results files. Running a practice participant and keeping a log book will be of great help in running the experiment smoothly.

Figure 2: Example of a Log Book Page

Participant number: 7 date: 7/11/06

List number: 4_1 (=IV)

Time in: __9.30__ Timeout:__ 10.15___

Handed: Right/~~Left~~ R

Experimental files: Notes:

Practice	Ran twice
Exp file/edat name	LD41_7_1

Backup eDat?	Merged?
Y	Y

Comments: Partic. has her face very close to the screen. Says she forgot her glasses_____

Chapter 5 Data Processing and Statistical Analysis

This chapter is devoted to the final stage of conducting an experiment: performing a statistical test on your data to see whether your predictions are borne out. We will first repeat and extend the discussion of Chapter 1 of why statistics are needed, and how analysis can be made simple if you first clearly specify your design and predictions in terms of a factorial design (Section 5.1.2). Next, we will deal with the more practical parts of data processing and statistical analysis (Sections 5.2 to 5.4). The discussion will be limited to t-tests and ANOVAs, since these are the types of analysis typically used with RT data. Finally we will deal with interpreting the results and assessing the experimental hypotheses (Sections 5.5 and 5.6).

5.1 Statistical analysis

5.1.1 Why Statistics

Why do we need statistics? We already have discussed this issue in Chapter 1, but let us repeat the major points. Recall AB's Hypothesis 1":

Hypothesis 1''

> Given an ambiguity, people choose one reading immediately (on the basis of strategies, frequency…). As a consequence, the chosen reading (in this case, NP coordination) will be easier to understand than the alternative which is not initially chosen (in this case, S-coordination). This will show up as an increase in difficulty when information disambiguates towards the alternative reading (due to reanalysis or a failure to comprehend the structure).

According to this hypothesis, object relative clauses should take longer to read or to make a grammaticality decision on than subject relative clauses. Suppose that AB ran the experiment with these two conditions and that the mean RTs came out as follows:

NP coordination:	470.5 msec
S coordination:	539.5 msec.

This means that the object relative clauses are reacted to 69.5 msec more slowly than subject relatives. This is a fairly large difference. On the face of it, AB's prediction is borne out. But what if the condition means were 470.5 and 503 msec for the NP- and S-coordinations, respectively, that is, a difference of 32.5 msec? Would her prediction still be borne out? What about a difference of 5 msec? The difference in means may be a mere coincidence, and may not have occurred if AB used different sentences or a different group of participants. Obviously, AB does not have the time and money to test all people in the world and all possible sentences containing coordination ambiguities to see whether the difference she found reflects a real difference between the processing of NP and S coordinations. This is why statistics are needed.

Statistical tests were developed in order to still say something about the reliability of a difference based on a sample, in our case, a selection of possible sentences and a selection of possible participants. In other words, statistics allow you to ***generalize*** your results ***over participants*** and ***items***. A statistical test will tell you how likely it is that the difference found between conditions in an experiment is a real difference. This likelihood is expressed as a ***p-value***, which represents the **probability** that the conditions do **NOT differ** in reality.

Imagine that in reality there is no difference between the conditions, for instance, in reality people do not process NP coordinations differently from S-coordinations. In that case, all the differences we see between the conditions in our experiment are due to other factors. If we do the same experiment a number of times, each time obtaining a set (sample) of response times for the conditions, the mean reaction time values for the conditions will of course differ for each experiment. However, since there is no difference between the conditions in reality, the mean values for the conditions are more likely to be close to each other than very different from each other; after all, the chance of obtaining only extreme values in your samples is rather minimal. On the basis of the variability observed in the data, we can see how often two samples will be extremely different from each other, even though in reality, the conditions do not differ. You can understand a p-value of .726 as meaning that you can expect to get this big of a difference between the means 726 times in a thousand if you pick random samples out of a set of response times (which do *not* in fact systematically differ), given the variance in the entire set of data. It is very unlikely that two random samples will be picked that have a very low p-value (e.g., p = .001, a difference this large will be seen once in a 1000 random choices). This means that if p is very small, the conditions are very likely to actually come from **different** distributions, that is, it is likely that the conditions differ in reality. Conventionally, a difference is considered to be ***significant*** if the p-value is smaller than .05. [2]

How can we determine the chance that a given result occurs by accident? The p-value is determined by (i) the difference in means of the conditions, (ii) the variability (variance or standard deviation) of the responses within each condition and (iii) the number of participant/items involved. The larger the difference in means and the smaller the variance, the *smaller* the chance that the conditions do not differ in reality. The number of participants and sentences in each condition also matters: the more data points you have in your experiment and the better estimate you can make of the means and the variance. If you pick two hundred responses you have a better idea of what data looks like in reality, than if you pick only two responses. In a statistical test, the number of independent participants/items on which a test is based is expressed as part of the ***degrees of freedom***, abbreviated as ***df***. The degrees of freedom, together with the statistical value (t-value or F-value) determine the probability of the difference between the conditions. If, for example, you consult a table of F-values in a statistics book to determine whether the value F=4.89 is statistically significant, you will find that it has a p-value of <.25 for 2 participants, <.1 for 4 participants, <.05 for 11 participants, and that you need to have a sample of more than 120 to have it significant at a level of <.025.

---

[2] This criterion, or *alpha*, depends on the nature of the hypothesis tested (sometimes it better to have false positives than undetected problems) and the number of tests conducted (the more, the smaller the alpha).

As we said above, you would like to generalize over participants as well as items. Both participants and items are random factors: you are not testing every person and every possible item, but you just collect data from only a few people, say 40, and only a few sentences, say 20. As an experimenter, you want to be able to say more than that these forty participants behaved in such and such a way, while reading these twenty particular sentences. Instead, you want to say that something about people *in general* processing these sentences, and about these types of items *in general* being processed. This means that each statistical test has to be done twice: once using the ***participant means*** for each condition to generalize over participants, once using ***item means*** for each condition to generalize over items. (We will illustrate how these means are calculated below in Section 5.3). If the participant analysis produces a significant result, then we can expect other participants **reading the same sentences** to behave in the same way. If the item analysis produces a significant result, we can expect other items to be responded to in the same way **by the same participants**. Only if the p-value is smaller than .05 in the participant **as well as** the items analysis can the differences be regarded as significant. Only in this case are the results reliable enough to say something about participants AND sentences in general.

5.1.2 Specifying Within and Between Participant/Item Factors

Carrying out a statistical analysis is much more straightforward if you can clearly indicate what you are comparing, and what pattern of results you expect. In Chapter 1, we already discussed the statistical design in terms of factors and levels. In this section, we will show how a design can be more precisely defined by specifying ***between*** and ***within participant*** and ***item factors***.

Let us illustrate these terms with AB's more complex experiment with four conditions: (A) NP-coordination, 'and' followed by an inanimate NP; (B) S-coordination, 'and' followed by an inanimate NP; (C) NP- coordination, 'and' followed by an animate NP; (D) S-coordination, 'and' followed by an animate NP. For convenience, one of the sentence sets in her materials is given below:

Item #3

a. The hairdresser recognized the brush and the comb at first glance. [+NP, inanim]
b. The hairdresser recognized the brush and the comb was also familiar. [+S, inanim]
c. The hairdresser recognized the brush and the doctor at first glance. [+NP, anim]
d. The hairdresser recognized the brush and the doctor was also familiar. [+S, anim]

Sentence Sets	Item Groups	Participant List I	List II	List III	List IV
1-4	I	A	B	C	D
5-8	Ii	B	C	D	A

9-12	Iii	C	D	A	B
13-16	Iv	D	A	B	C

Recall from Chapter 3, that if you have four conditions, you will also have four participant lists and four item groups. AB had constructed 16 sentence sets. She has assigned her sentence sets to item groups (i-iv), participant lists (I-IV) and conditions (A-D) in the following way:

Let us start with the within and between participant factors in AB's design. In this experiment, no participant sees more than one participant list. Participants can thus be split into groups on the basis of which list they have read.

- List, therefore, is a *between* participant factor. Because AB had four lists, the factor List is a factor with four ***levels***.

Each of AB's participants saw NP- as well as S coordinations.

- Type of coordination, therefore, is a *within*-participants factor with two levels: NP- and S-coordination.

Each participant also saw sentences containing an animate NPs following 'and', and sentences containing an inanimate NP following 'end'.

- Animacy, therefore, is another within-participants factor.

This factor also has two levels: animate and inanimate .

Now let us turn to the within- and between-item factors. Recall that sentence sets are divided into item groups. No sentence set appears in more than one item group, so

- Item group is a *between item factor*, with as many levels as there are item groups (four in AB's experiment).

Each sentence set has four members: one version of a sentence for each condition. The conditions in AB's experiment are distinguished by the type of coordination that the sentence contains, and by the animacy of the NP following 'and'. Therefre, in AB's experiment,

- Type of coordination and Animacy are also *within item factors,*

each with two levels: NP and S coordination; and animate or inanimate NP, respectively.

5.1.2.1 Why include list and group as between participant/items factors?

Your statistical design should always include the between factor Participant List in the participant analysis and the between factor Item Group in the item analysis. The most important

reason to do so is that this factor accounts for some of the potential variance in your data, so that you are more likely to achieve a significant result on the effects of interest if these factors are included; we will not take the time to discuss the mathematical basis of this point in detail. Because it accounts for variance, it also serves as a check to see whether the item groups were properly matched and whether the participants in the various lists differed concerning speed of response and sensitivity to the experimental conditions. If your materials are properly matched, the main effect of Item Group should not be significant. If your participants do not differ in their ways of responding, there should be no main effect of Participant List. In addition, the lack of interactions between (i) Participant List and Condition (e.g. Type of Coordination) and (ii) Item Group and Condition will also tell you that there is no difference between your materials and participants, respectively.

This may sound somewhat counter-intuitive: Why does an interaction between Item Group and Condition have anything to do with differences across participant lists? To see this, suppose that you ran an experiment with two conditions: NP coordination and S-coordination. Two conditions means two participant lists (I and II) and two item groups (i and ii). Suppose the sentences and conditions were assigned to the lists as follows:

	List I	List II
Item Group i	NP-coordin.	S-coordin.
Item Group ii	S-coordin.	NP-coordin.

Suppose that coincidentally, participants in List I were relatively slow readers, whereas participants in List II were relatively fast. The resulting means per condition and sentence group are given below:

	NP-coordination	S-coordination
Group i	448 (slow participants, List I)	461 (fast participants, List II)
Group ii	443 (fast participants, List II)	608 (slow participants, List I)

As result of the differences in speed between the participant groups, the S-coordinations in Item Group i are read almost as fast as the NP-coordinations. In Group ii, however, NP coordinations are read much faster than S-coordinations. This leads to a Condition (i.e. Coordination Type) by Item Group interaction, which is due to fast participants being assigned to List I. So, although it might appear somewhat counter-intuitive, an Item Group by Condition effect may be due to your participants rather than your materials. Similarly, a Condition by List interaction may be due to your items and not to your participants if one Item Group contains one or more items that are reacted to substantially faster than other sentences, and another group contains one or more items that are read substantially slower (just reverse the labels in the table above to see this for yourself.

5.1.3 Statistical Analysis: Some Important Points

To summarize, since your hypotheses and predictions concern language processing in general, you must test whether the results of your experiment can be generalized across participants and items. This means you have to do two statistical analyses: one on the participant means, and one on the item means.

For each analysis you have to specify the factors in your design. Your experimental manipulations will most often be formulated as within-participant or within-item factors. However, Participant List and Item groups should always be included as between-participants and between-items factors in the analysis by participants and items, respectively. This serves as a check to see whether your materials are properly matched and the participants in the various lists do not differ concerning the way of responding. In addition, including these between factors will absorb some of the noise due to these factors, increasing the chance of obtaining a significant result if there is a real difference between the conditions.

In the rest of this chapter we will discuss how to conduct a statistical analysis on RT data. First, the data must be screened for results that are too long or too short to be genuine responses. These RTs will only increase the variance in your data and decrease the chance of finding significant differences. In Section 5.2.1 we will show how to manage such deviant RTs by cutting off and trimming the data. Next, the mean RTs per condition need to be calculated for each participant (participant means) and item (item means) (Section 5.2.2). Third, you need to feed these means into SPSS, or another statistical software package, and have this program perform an analysis. It is not our purpose here to teach you statistics. We will therefore NOT deal with the underlying statistical theory and assumptions or give an extensive introduction to SPSS.

5.2 'Cleaning' Your Data

5.2.1 Treating Outliers as Missing Data

5.2.1.1 What are outliers and how do they affect your data?

Let's take a concrete example to illustrate the several stages of data processing. Suppose experimenter AB ran her experiment contrasting NP and S coordination using 6 participants. She has collected grammaticality decision responses at the end of the sentence only. After she has merged het data files and deleted some irrelevant columns, her data look as follows.

Part.nr	list	itemnr	itmgr	condition	RT
1	1	1	1	1	450
1	1	2	1	1	390
1	1	3	2	2	467
1	1	4	2	2	654
2	2	1	1	2	30
2	2	2	1	2	542
2	2	3	2	1	334
2	2	4	2	1	432
3	1	1	1	1	421

3	1	2	1	1	357
3	1	3	2	2	497
3	1	4	2	2	493
4	2	1	1	2	550
4	2	2	1	2	549
4	2	3	2	1	467
4	2	4	2	1	475
5	1	1	1	1	478
5	1	2	1	1	342
5	1	3	2	2	446
5	1	4	2	2	547
6	2	1	1	2	534
6	2	2	1	2	495
6	2	3	2	1	979
6	2	4	2	1	479

All of her participants made no errors in the response, so she decides to look at the RTs only. The numbers in the rightmost column are the RTs in msec. For each response you have to know which participant made it and which list this participant saw, which item it was a response to, which item group this sentence was a member of, and lastly the condition in which the item appeared. This information is given in the first five columns: the first number on a row is the participant number, the second the list number, the third the item number, the fourth the item group number, and the fifth the condition, '1' in this column stands for 'NP-coordination', '2' for 'S-coordination'. In this way we know that participant number 1 who read list number 1 read sentence number 1 (which is a member of item group 1) as a NP-coordination and responded after 450 milliseconds.

Of course, you are interested in the results per condition, so let us calculate the means nd standard deviation for each condition per condition across all six participants per condition. We then get the following values.

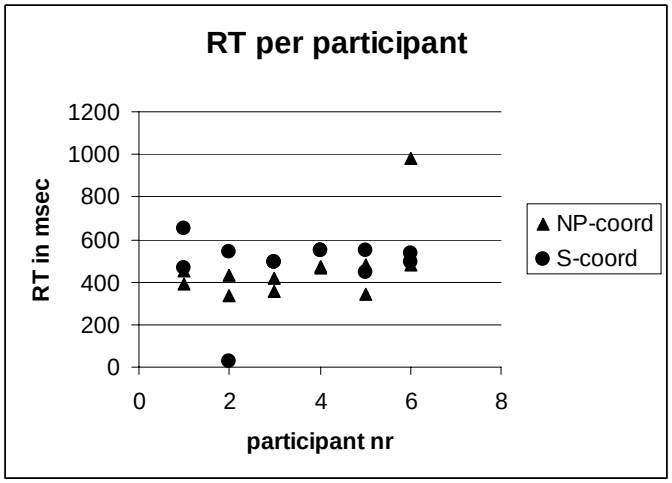
	Condition	
	1 (NP coord.)	2 (S coordin.)
mean:	467	484
stdev:	170	152

If you look at the means, you'll say that Condition 1 is about the same as Condition 2: the reaction times to items in Condition 2 is only 16.67 msec slower than the reaction times in Condition 1. This suggests that participants read S-coordinations at about the same speed as NP-coordinations, contrary to the predictions of the hypothesis developed in Chapter 1. The variance in the data supports this, since the standard deviation is many times larger than 16 msec.

But let us look at the data more carefully by plotting the reaction times per participant. Figure 1 shows these data: the dots represent the RTs to NP coordinations, the triangles the RTs to S-

coordinations. Whereas most responses can be found around 300 to 550 msec, two responses clearly are not in this range: the 30 msec RT in Condition 2 (S-coord.) of participant 2 and the 979 msec RT of participant 6 in Condition 1 (NP-coordin). Data like these, that clearly fall outside the range of normal responses, are called ***outliers***. Are these data real reflections of decision times? It is very unlikely that the short response is. In 30 msec you have not even seen a word clearly: it takes about 60 msec for the visual information to be identified clearly enough for a word to be recognized. So clearly this is NOT a decision that the sentence is grammatical, but an error of some sort. What about the 979 msec? Here some other processing appears to have taken place in addition to the processes needed to make a grammaticality decision, since the RT is nearly twice as long as most other responses! This additional processing is not what AB meant to investigate. These outlying responses clearly influence your means much more than you want them to. If we look at the means with these data points left out, the means and the standard deviations alter dramatically.

Figure 1



In AB's case, the two outliers make the difference between the two conditions seem much smaller than it actually is. In addition, they render the variance in RTs **quite** large, which negatively influences the likelihood of finding a difference if there is any. As a rule of thumb, a difference which is larger than the standard deviation produces a significant difference. Without the outliers, the difference between the means of the NP-coordination and S-coordination conditions is substantially larger than the standard deviation. So, it is best to minimize the occurrence and effect of such data points. How can we do that?

5.2.1.2 Absolute cutoffs

In psycholinguistic research, most experimenters choose a maximum and a minimum value within which responses are considered valid. These values are called ***absolute cut-off*** values. The minimum and maximum values can vary depending on the kind of experiment and kind of

task. For word-by-word self-paced reading experiments, a minimum of 90 and a maximum of 3000 msec are reasonable as absolute lower and upper cutoff values.[3] When a response is smaller than the minimum cutoff-value, it is very likely that the participant was not paying attention or accidentally pressed the button on the button box without having read the stimulus word. If a response is larger than the maximum, the participant probably was distracted by something, or the button box did not work properly. These very long or very short responses will not tell you anything about the 'real' processing you are interested in. That is why these outliers are often treated as ***missing data***. That is, they are omitted from the statistical analysis.

You can replace such outlying data in the result file by '99999', a '0' or another symbol (cf. Appendix 5) and let the statistical analysis programs you are using know that this symbol stands for 'missing'. Alternatively you can define criteria for absolute cutoffs in the statistical software package itself.

5.2.1.3 Dropping invalid responses

There is one other kind of data trimming that you might do, namely dropping all data points of a particular participant or a sentence. If you notice that one participant did not understand the task, you might throw him or her out (of the analysis, not the room). Or you might notice that one participant was excessively variable. One of the authors of this text, for instance, had a participant once who read mostly at the speed of 250 msec per words but every once in a while would stare at words for 10 seconds, probably trying to figure out what the experiment was about. Another of her participants went to sleep during the experiment! In the last half of the experiment all the response times were the maximal time after which the program goes on without a response. In these cases, you might want to consider that participant as undesirable noise in the experiment and no include his/her data.

Similarly, if responses to a particular sentence are too variable or strange, one can decide not to include this sentence in further analysis. Suppose that AB had forgotten to correct the typo in the following sentences:

> The hairdresser noticed the doctor and the nuser in the store.
> The hairdresser noticed the doctor and the nuser was also familiar
> .

She found that many participants found both of these sentences ungrammatical because of the typo. This sentence should be excluded from the analysis in this case. On the other hand, it would have been even better if AB had checked her materials better. Remember that the more sentences and participants that CAN be included in the analysis, the more data points you have and the smaller the variance.

---

[3] These criteria are rather arbitrary, and may vary depending on the technique, items and participants. To motivate these criteria, use the SPSS Explore tool to determine extreme outliers.

Another situation under which you might consider excluding data is when a participant has answered the trial incorrectly. Say that you record sentence readings times and performance on a following comprehension question. If the participant did not answer the question correctly, you could argue that s/he did not read the sentence carefully enough and that you are not interested in the participant's reading times for the sentence.

5.2.2 Trimming Data on the Basis of Variance

Suppose that AB used the absolute boundaries of 90 and 3000 msec as lower and upper cutoff values. This means she got rid of the 30 msec RT of participant #2, but not of the 979 msec RT of participant 6. Without the 30 msec, but still including the long RT, the new condition means are 467ms for the NP coordination and 525ms for the S-coordination. This is still not a tremendous difference, and the variance in Condition 1 (NP-coordination) is still quite large. Of course, AB could have chosen for an upper boundary of 1500, 900 or even 700msec. But the lower the upper cutoff, the fewer responses remain on which to include in the analysis, which means you are back at where you started: a large variance, and a smaller chance that a real difference between the conditions will be detected. In addition, it is hard to give a reasonable argument for why you have chosen a particular absolute cutoff. So, arbitrarily decreasing the upper cutoff is not a very reasonable way to eliminate very long response times.

A different approach is to calculate *relative cutoffs*. We argued that 979 msec was an unusually long RT because of the distribution of the data (cf. Figure 1). One way to express that distribution numerically is in terms of the standard deviations or variance. The standard deviation of the entire data set in our example is 129.28 (excluding missing data, such as the 30 msec data point which is outside the absolute cutoff). The mean of the entire data set is 494.7 msec. In a normal data distribution, 95% of the data lies within two standard deviations of the mean. We could use this to set a cutoff: data that are a number of standard deviations away from the mean (say 2.5 standard deviations) could be rejected.

However, this criterion is not very useful for relatively slow or relatively fast participants that do show the general pattern of results. Suppose AB ran several more participants, among others participants 7 and 9:

7	1	1	1	1	746
7	1	2	1	1	797
7	1	3	2	2	832
7	1	4	2	2	923
9	2	1	1	2	279
9	2	2	1	2	285
9	2	3	2	1	245
9	2	4	2	1	264

Suppose we calculated a cutoff based on the entire data set, as mean $\pm$ 2.5*SD, and saw that two of participant 7's responses (the two responses in the S-coordination condition) would actually be rejected. Is that reasonable? No; they are within the normal range for Participant 7. The fastest response of Participant 9 would also be rejected, although it appears to be in a normal

range for this participant as well. The data of these two participants is quite consistent, but we have one very fast participant and one very slow participant. These slow or fast responses are not outliers for these particular participants.

Given that (i) it is not a good idea to eliminate data if you are not sure that there is something wrong with it; (ii) that the distribution of the data can be used to estimate if a long or fast response is real; but that (iii) participants vary in speed, experimenters have come up with the following strategy for dealing with these questionable outliers: data that are long enough to possibly influence the results, but too short to be considered genuine responses are ***trimmed***, rather than regarded as missing data. In this way, long responses that could be genuine, and not due to e.g. a participant falling asleep, will still be included in the analysis; however, they will not have a disproportionately large influence on condition mean or variance, so that the statistics will only come out if the same pattern is also seen with other sentences and other participants.

The trimming procedure mostly used is the following.[4] First, extremely long or short responses are excluded by applying an absolute cutoff procedure. Next, for each participant, the mean response time is computed by adding all non-missing responses and dividing this total by the number of non-missing responses. Next, the standard deviation for this participant is computed. The ***relative upper cutoff*** for a participant is his/her mean reaction time plus twice or 2.5 times the standard deviation; the ***relative lower cutoff*** is the mean minus twice or 2.5 times the standard deviation. All response times larger than the upper cutoff are then replaced by the value of the upper cutoff; all response times faster than the lower cutoff are replaced by the value of the lower cutoff. This procedure is executed for each participant separately rather than on the total group data to avoid the problem just mentioned about very fast and very slow participants.

To illustrate this procedure, suppose we ran a participant, participant number 6, in an experiment with two conditions and eight sentences in each condition. The data are displayed below. The mean of all participant 6's RTs is 510.19 msec (we round off to 2 digits after the decimal). The standard deviation is 136.52. That means that the lower relative cutoff is 510.19 – (2.5*136.52)= 168.9 msec. The upper relative cutoff is: 510.19 + (2.5*136.52)= 851.48 msec. Nothing lies beyond the lower cutoff. One data point, 979 msec, is higher than the upper cutoff, so we will set it to 851 msec (rounded off to the closest decimal).

6	2	9	2	1	434
6	2	10	2	1	495
6	2	11	2	1	979
6	2	12	2	1	479

[4] There are a number of other methods. You may treat data longer(shorter) than the upper (lower) relative cutoff as missing data. This would give the overall mean reported earlier in the text, rather than the relatively high mean reported in the main text. Other researchers favor not using any cutoffs, as it is at odds with the idea that you are relying on a random data collection.

We have calculated our cutoffs using all the conditions pooled. That may be inappropriate if the conditions are sufficiently different. If you discover that the outliers are not equally distributed across the conditions, you may need to calculate the cutoff criteria separately for each condition.

6	2	13	2	1	450
6	2	14	2	1	467
6	2	15	2	1	390
6	2	16	2	1	354
6	2	1	1	2	547
6	2	2	1	2	554
6	2	3	1	2	493
6	2	4	1	2	482
6	2	5	1	2	477
6	2	6	1	2	494
6	2	7	1	2	506
6	2	8	1	2	562

Now, let's return to AB's data. We cannot demonstrate the relative cutoff procedure with the data of the six participants we gave at the beginning of this section: two sentences per condition per participant still yields a too large a standard deviation for the 979 msec to be reduced. (This shows that more that two sentences per condition should be used in a real experiment!). But suppose AB used eight sentences per condition illustrated and that the long response time lies above the relative upper cutoff for Participant 6. After trimming her data according to the procedure above, using both absolute and relative cutoffs, the means are as follows:

	NP coordination	S coordination
mean	456	526
stdev	135	56

Now the difference in mean RTs between NP and S-coordinations is much larger, and the standard deviations within the conditions much smaller than was the case before the data trimming.

5.3 Computing Participant and Item Condition Means

As we said in Chapter 1 and 5.1, by doing statistics, you typically wish to generalize over participants as well as over items. This means that each statistical test has to be done twice: once using the ***participant means*** for each condition to generalize over participants (in an ANOVA this is called the ***F1*** analysis), once using ***item means*** for each condition to generalize over items (in an ANOVA this is called the ***F2*** analysis). Let us illustrate these terms.

First consider the participant by participants. We will use the data reported for our participant 6 above as an illustration. The participant means for participant 6 are calculated by adding all of this participant's responses in one and the same condition and dividing this total by the number of non-missing responses in that condition for that participant. Participant 6 has no data outside the absolute cutoffs or lower relative cutoff. One RT has been set to the relative higher cutoff: that is, the 979 msec reaction time was replaced by 851 msec. The mean for Condition 1 (NP coordinations) for this participant is (434 + 495 + 851 +…+ 354)/8)= 490. The mean for

Condition 2 is calculated by the same procedure. The output of the participant means for participant 6 and some other hypothetical participants can be represented like this:

Part	list	cond1	cond2
6	2	490	514
8	2	520	572
9	2	463	434
10	1	548	568
11	1	651	796

The first column represents the participant number, the second this participant's list number, the third the mean RTs in Condition 1 (NP coordination), and the final column the RTs in Condition 2 (S coordination).

It is important to maintain all relevant data in the file. The original file contained a number(column) for the participant, list, item number, item group and for the condition. Since the items are averaged, and the condition is indicated by the relative order, Item group and item number are necessary anymore. Item group is predictable from List. Note that participant list is still represented. This is needed to include the factor Participant List in the statistical analysis.

Item means are obtained in the same way, except that for each sentence, the mean of the responses in each condition is computed across all the participants. For instance, given AB's data, the item data (simply sorted by item) is the following.

partc	list	**item**	group	cond	RT	
1	1	1	1	1	450	
3	1	1	1	1	421	
5	1	1	1	1	478	
7	1	1	1	1	746	
8	2	1	1	2	279	
2	2	1	1	2	30	→ missing
4	2	1	1	2	550	
6	2	1	1	2	534	
1	1	2	1	1	390	
3	1	2	1	1	357	
5	1	2	1	1	342	
7	1	2	1	1	797	
2	2	2	1	2	542	
4	2	2	1	2	549	
6	2	2	1	2	495	
8	2	2	1	2	285	
2	2	3	2	1	334	
4	2	3	2	1	467	
6	2	3	2	1	979	→ replaced by 851
8	2	3	2	1	245	
1	1	3	2	2	467	

3	1	3	2	2	497
5	1	3	2	2	446
7	1	3	2	2	832
2	2	4	2	1	432
6	2	4	2	1	479
4	2	4	2	1	475
8	2	4	2	1	264
1	1	4	2	2	654
3	1	4	2	2	493
5	1	4	2	2	547
7	1	4	2	2	923
…					

The means for sentence number 1 are calculated as follows: Condition 1: (450+ 421+478+ 746)/4= 523.75 ; Condition 2: (279 + 550 + 534)/3= 454.33. Note that one data point is missing in the latter case! The item mean data for these four sentence sets are thus:

Item	group	cond1	cond2
1	1	524	454
2	1	472	468
3	2	474	561
4	2	413	654

The last two columns represent the mean RTs in Condition 1 (NP coordination) and Condition 2 (S coordination), respectively. The first column represents the item number. The second column represents the item group the sentence belongs to. You will need information about items group to include Group as a between items-factor in your analysis. Participant number is no longer relevant as the data are collapsed over participants. Participant List is predictable from Item group and condition. In practice, you do not have to calculate means by hand: files containing participant and item means can be created in SPSS. The next step is to conduct a statistical analysis on these participant and item means. Below we will show how to analyze RT data using the SPSS/PC+ software, using AB's more complex two factorial design as an illustration.

We will restrict the discussion to a General Linear Model repeated measures analysis. We have chosen for this approach because the SPSS syntax can easily be adapted to more complex experimental designs. Bear in mind, that the statistical approach you choose has consequences for the way the data file is formatted and read by SPSS. So, consult an expert if you decide to conduct the analysis in a different way than described below.

5.4 SPSS

SPSS (Statistical Package for the Social Sciences) can import a series of numbers, that is, your file with participant or item means as input, but you need to tell it what the numbers are, that is specify what each column of numbers means. Once you have done this and saved the data as an SPSS data file (*.sav), you can test the data statistically in various ways. SPSS reports the results

of each analysis in an output file, which you can save for later use. In this output file you can find whether the effects you are interested in are significant.

5.4.1 Data File format

In order to be able to do a GLM analysis statistics, the data need to have a certain format and imported into SPSS.

This is what the format should look like for a file containing the means for each participant. This experiment has four conditions, as in AB's experiment contrasting the factors Coordination type and Animacy.

Partic	List	NPan	San	NPinan	Sinan
1	1	461	789	478	646
2	2	567	890	639	987
3	3	432	937	480	789
4	4	578	486	307	778
5	1	245	368	336	564
6	2	318	431	302	357
7	3	658	698	352	599
8	4	769	777	543	659
9	1	846	901	753	827
10	2	567	489	371	598
11	3	478	498	447	955
12	4	389	447	787	680
13	1	976	948	653	824
14	2	689	543	782	334
15	3	571	389	434	578
16	4	518	324	349	448

In the first two columns, the participant number and list number is given. The third through sixth columns contain the reaction times to the four experimental conditions. The lab exercise will walk you through the steps of importing a file into SPSS, dealing with missing data, and restructuring it to the format above.

Once the data are correctly read into SPSS, the file can be saved as an SPSS data file via FILE menu on the toolbar. Once saved, it can be accessed again without having to specify the variables, or missing data, since this information is saved with the file.

5.4.2 Calculating Means

An overview of the means per condition will make it a lot easier for you to interpret your results and any differences between the conditions. You can have SPSS make a print out of the means per condition by clicking ANALYSIS, followed by DESCRIPTIVES, then choosing the variables, for example NPan, San, NPinan, and Sinan, for which you want means and standard deviations. Other statistical descriptions can also be selected, through the Options menu. For AB's data, the output would look like this.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
NPan	16	245.00	976.00	566.3750	190.85313
San	16	324.00	948.00	619.6875	222.90318
NPinan	16	302.00	787.00	500.8125	171.89424
Sinan	16	334.00	987.00	663.9375	190.82050
Valid N (listwise)	16				

Y also may want to know how the reactions on the various lists compare with each other. You can get this information by selecting ANALYSIS, then COMPARE MEANS, followed by MEANS. At this point you specify the dependent variables, for example NPan, San, NPinan, and Sinan. This version wants to break the data down into subgroups, for example, the different lists (which can certainly be interesting information). Thus you would specify list on the INDEPENDENT LIST of variables to break the data up into. For each variable (NPan, San, NPinan, and Sinan) the number of instances, mean and standard deviation will be reported for each ***level*** of the factor list (each participant list: here 1.00 to 4.00) considered separately and for all participant lists taken together (Total). The output of this command would look like:

Report

List		NPan	San	NPinan	Sinan
1.00	Mean	632.0000	751.5000	555.0000	715.2500
	N	4	4	4	4
	Std. Deviation	338.20211	264.22276	185.02072	131.63934
2.00	Mean	535.2500	588.2500	523.5000	569.0000
	N	4	4	4	4
	Std. Deviation	155.83405	206.29974	225.44844	303.16882
3.00	Mean	534.7500	630.5000	428.2500	730.2500
	N	4	4	4	4
	Std. Deviation	100.47014	241.09265	54.39593	177.36098
4.00	Mean	563.5000	508.5000	496.5000	641.2500
	N	4	4	4	4
	Std. Deviation	158.07277	191.85151	219.25556	138.87974
Total	Mean	566.3750	619.6875	500.8125	663.9375
	N	16	16	16	16
	Std. Deviation	190.85313	222.90318	171.89424	190.82050

At the bottom (Total) the mean and standard deviation are listed for all 16 participants for each of the four variables specified. In the rows above, the data is broken down by participant list. It is interesting to look at the values for the different lists, because the participants in a particular list could be slower or faster than in another list. In the data above, participants in List 1 were slower than participants in the other lists. Including List as a between-participant factor can help dissipate some of this variability, as can item group (as a between-item factor), which is why we include them as factors in the analysis discussed below.

5.4.3 A Repeated Measures analysis

We will now deal with the actual repeated measures analysis. First, you need to specify the factors in the design. In AB's study each participant sees items and gives responses in each of the four conditions. Each participant is therefore a control for him/herself. This that means that we are dealing with a ***repeated measures*** design. We are only dealing with the participant means for now, so only the between- and within-**participants** factors are relevant. In AB's case, List is a four-leveled between-participants factor, Coordination Type (or 'coor' for short) and Animacy ('anim') are both two-level within participants factors. Between them, they determine four dependent variables.

Select ANALYSE, then GENERAL LINEAR MODEL, followed by REPEATED MEASURES. At this point you need to name the factors and specify how many levels they each have. In this case, that would be for example anim – 2 and coord – 2 (clicking ADD after both name and levels have been specified). At this point you define these variables; that is, you need to tell the program which variables express each combination of levels of the factors.

The factors specified, *anim(2)*and *coord(2)* in the example, tell SPSS to take the variables names which you specify and separate them into four groups. SPSS reminds you of this by giving an abbreviation of the levels. So if you add the four variables *NPan San NPinan Sinan* in the following order:

Variable	levels
NPan	(1,1)
San	(1,2)
NPinan	(2,1)
Sinan	(2,2)

The first two will be regarded as being level 1 of the factor Anim and the second two will be regarded as being the second level of this factor. The first and the third will be regarded as being level 1 of the factor Coord factor (NP-coordination), and the second and fourth will be regarded as being the second level of this factor (S-coordination).

You also need to specify any between participant factors. In this case, *list* is a between participants group (since a participant only received one list, not all of them). Select this variable and click ADD at the box where between- factors are specified.

Note that the analysis described above is only the analysis on the participant means (F1-analysis, or analysis by participants). In order to generalize over items, a second analysis (F2-analysis, or analysis by items) must be conducted on the item means. Such an analysis is conducted in exactly the same way, except that, of course, item means are used. Anim and Coord are specified as within-item factors, with Item group as a between-items factor.

5.5 Reading the Output of a GLM analysis

We will discuss the output of REPEATED MEASURES analysis in some detail. The tables that are most important for the basic statistical results appear in the text below—use them to find and interpret the information you need in your own output.

First there are sections of the output which you should check to see that analysis has been carried out correctly. In the first part of the output, the within-factors are listed. Check whether the assignment of the variables to the factors is correct before proceeding.

Within-Subjects Factors

Measure: MEASURE_1

anim	coord	Dependent Variable
1	1	NPan
	2	San
2	1	NPinan
	2	Sinan

The second table will give an overview of between-factors:

Between-Subjects Factors

		N
List	1.00	4
	2.00	4
	3.00	4
	4.00	4

Here we see that SPSS has read 16 cases (the data of 16 participants) spread over four lists.
Always check to see that the appropriate number of cases has been read, if not there is some error in your specification of the data.

5.5.1 Univariate ANOVA Output

SPSS reports two kinds of analysis after running a repeated measures analysis; one is an univariate ANOVA; and the other is a multivariate analysis. We will only deal with the univariate case, since this is the most common. We will first discuss the relatively simple case of

two-level factor univariate ANOVA results and then more than two level factors, using AB's data above.

In a 2x2 factorial design, such as AB's, you would like to know the significance of (a) the main effects of the factors of interest, e.g. main effects of Coord and Anim; and (b) the interaction of two or more factors, e.g. the interaction of Coord with Anim, or List with Coord. In journal articles, statistical results are generally reported in the following format: [F(*df1*,*df2*) = *F-value*, p = *p-value*], for instance: F(1,58)=4.50, p<.05. This is done for both participants and items analysis, F1 and F2, respectively. To obtain the df (degrees of freedom), F and P-values needed, find the table in the output with the title 'Tests of Within-Subjects Effects'.

Tests of Within-Subjects Effects

Measure: MEASURE_1

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
anim	Sphericity Assumed	1816.891	1	1816.891	.073	.791
	Greenhouse-Geisser	1816.891	1.000	1816.891	.073	.791
	Huynh-Feldt	1816.891	1.000	1816.891	.073	.791
	Lower-bound	1816.891	1.000	1816.891	.073	.791
anim * List	Sphericity Assumed	16338.297	3	5446.099	.220	.881
	Greenhouse-Geisser	16338.297	3.000	5446.099	.220	.881
	Huynh-Feldt	16338.297	3.000	5446.099	.220	.881
	Lower-bound	16338.297	3.000	5446.099	.220	.881
Error(anim)	Sphericity Assumed	297044.063	12	24753.672		
	Greenhouse-Geisser	297044.063	12.000	24753.672		
	Huynh-Feldt	297044.063	12.000	24753.672		
	Lower-bound	297044.063	12.000	24753.672		
coord	Sphericity Assumed	187380.766	1	187380.766	6.252	.028
	Greenhouse-Geisser	187380.766	1.000	187380.766	6.252	.028
	Huynh-Feldt	187380.766	1.000	187380.766	6.252	.028
	Lower-bound	187380.766	1.000	187380.766	6.252	.028
coord * List	Sphericity Assumed	66841.672	3	22280.557	.743	.547
	Greenhouse-Geisser	66841.672	3.000	22280.557	.743	.547
	Huynh-Feldt	66841.672	3.000	22280.557	.743	.547
	Lower-bound	66841.672	3.000	22280.557	.743	.547
Error(coord)	Sphericity Assumed	359638.813	12	29969.901		
	Greenhouse-Geisser	359638.813	12.000	29969.901		
	Huynh-Feldt	359638.813	12.000	29969.901		
	Lower-bound	359638.813	12.000	29969.901		
anim * coord	Sphericity Assumed	48235.141	1	48235.141	2.910	.114
	Greenhouse-Geisser	48235.141	1.000	48235.141	2.910	.114
	Huynh-Feldt	48235.141	1.000	48235.141	2.910	.114
	Lower-bound	48235.141	1.000	48235.141	2.910	.114

anim * coord * List	Sphericity Assumed	35920.797	3	11973.599	.722	.558
	Greenhouse-Geisser	35920.797	3.000	11973.599	.722	.558
	Huynh-Feldt	35920.797	3.000	11973.599	.722	.558
	Lower-bound	35920.797	3.000	11973.599	.722	.558
Error(anim*coord)	Sphericity Assumed	198893.313	12	16574.443		
	Greenhouse-Geisser	198893.313	12.000	16574.443		
	Huynh-Feldt	198893.313	12.000	16574.443		
	Lower-bound	198893.313	12.000	16574.443		

The column 'Source' contains the factors and their interaction. Main effects are represented by a single factor (e.g.' LIST', 'COORD' or 'ANIM'); interactions are represented by the name of two or more factors, separated by '*' (e.g. 'COORD * LIST'). You see that four rows of results values are given for each effect, but that they do not differ in the current results. We get back to this below. The column headed by 'DF' contains the degrees of freedom; 'F' contains the F-value, and 'Sig' the p-value.

The ***degrees of freedom*** or ***df*** is the number of values within the sample that can freely vary and still lead to the overall mean of the sample. Two dfs need to be reported; numerator and error df. The error df is related to the number of participants or items. When dealing with a single sample set, this df is the total number of participants (items) minus 1, because the last value is fixed by the mean. When dealing with several sample sets (between factors, e.g., the four participant lists), the total degrees of freedom is the total number of participants minus the number of sets (lists or groups). The numerator of the degrees of freedom expresses the number of levels in the factor. The df is the number of levels minus 1, since the value of the last level is fixed by the mean.

For the effect of COORD, the degrees of freedom are 1 and 12: 1 comes from the row containing COORD as factor (and equals the number of levels minus 1) and 12 comes from the 'error(COORD)' row. The error df is the number of participants (16) minus the number of levels of the between factor *list* (4).

The 'F'-value can be found in the column marked 'F'. For the effect of Coord, F= 6.25. The column 'Sig of F' contains the p-value of the test. As we pointed out in Section 1 of this chapter, a p-value represents the probability that the levels of the factor do not really differ. For the effect of Coord, p=.028. Conventionally, a difference is significant if the p-value is smaller than .05. You would report the effect of Coord as follows: $F(1,12)=6.25$, $p=.028$ (or: $p < .05$).

In AB's data the main effect of type of coordination is significant, but none of the other effects are. For example, there is no significant interaction of List with Coord (coord *List). This means that the participants in each list reacted to NP- and S-coordinations in the same way.

Main effects of between-participant factors

The table of within-effects discussed above only lists effects of the between-factor if it interacts with one or more of the within-factors. Sometimes you do want to know whether there is a main effect of a between-factors, e.g. whether there is a general difference between the participant lists. These data are found in the table with the following heading:

Tests of Between-Subjects Effects

Measure: MEASURE_1
Transformed Variable: Average

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	22105277.641	1	22105277.641	225.601	.000
List	130576.547	3	43525.516	.444	.726
Error	1175809.063	12	97984.089		

This table is structured in the same way as the table reporting the within effects. Is the effect of LIST significant?

Participants(Subjects) and Items Analyses

In psycholinguistic research, a difference is generally only regarded as significant if a factor has a p-value of less than .05 in both the analysis by participants and the analysis by items. Why is this? Since we test *means over a set of items* for participants, there could be a single, very consistent item set that caused the entire effect. By testing the items as well, we can see that this is not the case, but that the pattern is seen consistently across item sets as well as consistently over participants.

As you could read from the SPSS output of AB's analysis, the main effect of coordination type is significant in the analysis by participants. The effect can thus be generalized over participants. However, you do not know whether the effect can be generalized over items as well. To do this, you need to know whether the p-value of 'COORD' is smaller than .05 in the analysis using the item-means. This allows you to generalize to items.

5.5.2 Interpreting Significance

If a factor is significant, does that mean that your prediction is supported? First you should check the mean values of the levels involved to see whether the difference is in the predicted direction. A significant main effect can be caused by condition A being read faster than condition B, but also by condition B being read faster than condition A. So, if you predicted A to be faster than B, you should check the means in order to see whether this prediction is borne out.

Let us look at AB's data. For the analysis by participants, a significant main effect of Coordination type was found. But which is read faster: NP or S coordintions? To answer this

question you need to look at the overview of the means shown earlier. The mean reaction times for the NP coordination is 566 msec in the animate condition and 501 msec in the inanimate condition. This renders a mean RT of (566+501)/2 =533.5 msec. The mean RT for the S-coordinations is (620+664)/2=642 msec. So, S-coordinations take longer to process than NP coordinations, just as predicted!

5.5.3 Factors with more than two levels

Thus far we have been dealing with designs in which the within-participants or item factors contain only two levels. For example, the factor Coord contained the levels 'NP-coordination' and 'S-coordination'. Sometimes within factors will contain more levels. This has consequences the p-value you report. In a data set with more than two levels, some assumptions about the distribution of the data may no longer hold, which may skew the results. SPSS provides a test and a correction for this. Let us illustrate this and show you how to find the correct p-value.

Suppose AB reran her first experiment, in which only two conditions, NP and S coordinations, were tested, and looked at reading times on the last three words instead of investigating end of sentence responses. This means that not one but three reaction times per condition need to be analyzed. AB wants to know whether the effect of coordination type differs across word positions. She therefore runs a GLM repeated measures analysis with six variables per participant/item, specifying as within participant factors word position (POS) with 3 levels and COORD with 2 levels. The part of the univariate output, in which the within participant factors are reported, is given below.

Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
pos	Sphericity Assumed	30122.063	2	15061.031	2.147	.139
	Greenhouse-Geisser	30122.063	1.363	22100.968	2.147	.158
	Huynh-Feldt	30122.063	1.862	16176.733	2.147	.143
	Lower-bound	30122.063	1.000	30122.063	2.147	.169
pos * List	Sphericity Assumed	25819.354	6	4303.226	.613	.717
	Greenhouse-Geisser	25819.354	4.089	6314.671	.613	.662
	Huynh-Feldt	25819.354	5.586	4622.003	.613	.707
	Lower-bound	25819.354	3.000	8606.451	.613	.619
Error(pos)	Sphericity Assumed	168376.917	24	7015.705		
	Greenhouse-Geisser	168376.917	16.355	10295.037		
	Huynh-Feldt	168376.917	22.345	7535.419		
	Lower-bound	168376.917	12.000	14031.410		
coord	Sphericity Assumed	229614.844	1	229614.844	5.311	.040
	Greenhouse-Geisser	229614.844	1.000	229614.844	5.311	.040
	Huynh-Feldt	229614.844	1.000	229614.844	5.311	.040
	Lower-bound	229614.844	1.000	229614.844	5.311	.040
coord * List	Sphericity Assumed	69988.365	3	23329.455	.540	.664
	Greenhouse-Geisser	69988.365	3.000	23329.455	.540	.664
	Huynh-Feldt	69988.365	3.000	23329.455	.540	.664
	Lower-bound	69988.365	3.000	23329.455	.540	.664
Error(coord)	Sphericity Assumed	518782.292	12	43231.858		
	Greenhouse-Geisser	518782.292	12.000	43231.858		
	Huynh-Feldt	518782.292	12.000	43231.858		
	Lower-bound	518782.292	12.000	43231.858		
pos * coord	Sphericity Assumed	12925.563	2	6462.781	.784	.468
	Greenhouse-Geisser	12925.563	1.602	8067.018	.784	.444
	Huynh-Feldt	12925.563	2.000	6462.781	.784	.468
	Lower-bound	12925.563	1.000	12925.563	.784	.393
pos * coord * List	Sphericity Assumed	53011.354	6	8835.226	1.072	.406
	Greenhouse-Geisser	53011.354	4.807	11028.368	1.072	.405
	Huynh-Feldt	53011.354	6.000	8835.226	1.072	.406
	Lower-bound	53011.354	3.000	17670.451	1.072	.398
Error(pos*coord)	Sphericity Assumed	197834.083	24	8243.087		
	Greenhouse-Geisser	197834.083	19.227	10289.244		
	Huynh-Feldt	197834.083	24.000	8243.087		
	Lower-bound	197834.083	12.000	16486.174		

Measure: MEASURE_1

You can see that for the factors containing more than two levels, the p-values are different for the four rows of results. To determine which result to report, you need to look at the table listed above the previous table in the output:

Mauchly's Test of Sphericity(b)

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Epsilon(a)		
					Greenhouse-Geisser	Huynh-Feldt	Lower-bound
pos	.533	6.930	2	.031	.681	.931	.500
coord	1.000	.000	0	.	1.000	1.000	1.000
pos * coord	.752	3.139	2	.208	.801	1.000	.500

Measure: MEASURE_1
Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.
a May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.
b Design: Intercept+List
Within Subjects Design: pos+coord+pos*coord

In this table output, tests are given to see whether the assumptions underlying the ANOVA approach hold for the current set of data sample, in particular the assumption of sphericity (roughly: data are unrelated between the levels). First, find the rows corresponding to the factors or interactions of factors with more than two levels. See whether 'Sig' of the Chi-square mentioned is >.05 (n.s.).

- If it is >.05, then sphericity holds. You can just report the p-value for this effect that is listed in the row 'sphericity assumed' in the 'Test of within subjects effects' table.
- If it is <.05 (as for 'pos' and 'pos * coord' in the above), sphericity is violated. This means that the p-values should be adjusted. We recommend reporting the p-values listed as 'Greenhouse-Geisser' in the 'Test of within subjects effects'. Although strictly speaking the dfs have been adjusted to obtain this p-value, you can report the usual p-values listed in the 'sphericity assumed' rows. In the above example, the effect of position can be reported as: F(2,24)= 2.15, p=.158 (not significant).
- Do this for all factors or interactions of factors with more than two levels.

5.5.4 Summary: Reporting Statistical Results

In sum, the significance of the effects is reported in the following way. First you mention the relevant factor, e.g. `main effect of Coordination type' or `interaction of animacy and coordination type.' Next, you report the F-value of the analysis by participants, including the degrees of freedom: e.g. F1(1,12)=6.25, p <.05.

- The 1 after the 'F' indicates the analysis by participants.

- The first number between the brackets is the degrees of freedom of the factor being tested (DF at the row with the factor name in ANOVA);
- the second number is the degrees of freedom of the sample of participants/items, which is called the error term ('error ' in ANOVA).
- Next, you report the actual F-value for the comparison
- and then the p-value (or whether it is smaller than .05; .025, .01 or .001). If a factor contains more than two levels, first see whether sphericity is violated. If the sphericity test is significant p <.05, then report the p-values in the 'Greenhouse-Geisser' row. Otherwise, report the p-values in the 'Sphericity assumed' row.

So, if your p-value is .020, you can report: F1(1,12) =8.56, p <.025.

- The F1 values are immediately followed by the values obtained in the analysis by items (F2) for this factor.
- Do not forget to report the **direction** of the difference, that is, which condition is reacted to faster relative to which other condition(s)!

Without this *direction* of difference it means nothing to know that the difference is significant.

So, a complete report for the main effect of coordination type (from the first analysis discussed above may look like this:

`The main effect of coordination type was significant [F1(1,12)=6.25, p <.05; F2(1,28)=9.35, p <.025]; NP coordinations were responded to 108.5 msec faster on average than S-ccordinations.'

If, contrary to prediction, a factor or interaction is not significant, you report this in the following way.

- When the F-value is smaller than 1, (e.g. the COORD BY LIST interaction in AB's first set of results discussed above) the following will suffice: [F < 1, n.s.] where 'n.s.' stands for 'not significant'.
- If the F-value is larger than 1, but the p-value is larger than .05 (cf. The ANIM BY COORD interaction), you report the full F and p-value: [F1(1,12) = 2.91, p =.114; (or 'p >.1,' ...].

5.6 Interpreting the Results

Now that we are almost at the end of the experimental line, let us go back to where we started. The first step in designing an experiment was to formulate a research question and one or more possible answers (hypotheses) to this question. From these hypotheses, specific predictions were derived about the way participants would react to specific sorts of materials. Now that the experiment has been run and the data are analyzed, you need to answer the following question: Are your predictions borne out? And is (one of) your hypothesis supported?

In AB's case, one of the hypotheses, Hypothesis 1'' was that in cases of structural ambiguity, people will commit themselves to a certain analysis right away. This lead to the prediction that S-coordinations would be more difficult to parse than NP coordinations. AB has worked her way through designing the materials, conducting the experiment and doing statistics to see if there is a difference between these two conditions.

Is her prediction borne out? Say AB found that the effect of Coordination was not significant. Does this mean that her hypothesis is false? Not necessarily. There are a number of reasons why the difference between NP and S coordination could not have been statistically significant. For instance, the method she used may not have been sensitive enough, or the materials varied too much in frequency too detect an existing difference. If a difference is not significant you can therefore never conclude that there is NO difference, and that the hypothesis is false.

On the other hand, say AB found a significant difference between NP and S coordinations, for both participant and item analyses, and that the difference was in the predicted direction: S-coordinations took longer to read and make a grammaticality decision to than NP coordinations. In this case, her prediction is borne out. Is her hypothesis true? Not necessarily. Remember that the results tell you that the alternative hypothesis, according to which no differences between the conditions are expected, is false. You cannot say on the basis of the results that the Hypothesis 1'' is **true**. The results could be potentially explained by dozens of other hypotheses. For instance, recall CD's objection from Chapter 1, that according to some linguistic theories, S-coordinations are more complex than NP coordinations. Hence, regardless of ambiguity, S coordinations will take longer to process than NP coordinations. So, all you can say on the basis of a given experiment is that the results are *consistent* with the predictions of a given hypothesis and therefore *support* the hypothesis, but you can never determine that there is no other potential explanation.

To make this latter point even clearer, suppose that your hypothesis is that syntactically more complex sentences are harder to process than syntactically simple sentences. Assuming that S-coordinations are syntactically more complex than NP-coordinations, you predict that (1a) takes longer to read than (1b).

(1) a. The actress kissed the photographer and the journalist was very happy.
b. The actress kissed the photographer and the journalist on their cheek

Suppose that indeed this prediction is borne out. However this result does not necessarily imply that syntactically more complex sentences take longer to read than syntactically simple sentences: even if S-coordinations and NP-coordinations are syntactically different, the two sentence types might differ on more dimensions than syntactic complexity: NP coordinations are more frequent than S-coordinations; NP coordinations sentences impose less restrictions on discourse and intonation than S-coordinations, and so on. So, the fact that S-coordinations sentences are more difficult could be attributed to various other factors besides syntactic complexity.

On the basis of one experiment you therefore cannot say that your hypothesis is true. The only thing you can say is that your results are compatible with the hypothesis. Alternative explanations are always available. The best way to proceed is to conduct a series of experiments,

carefully designed to single out each factor that may possibly contribute to the differences found between the conditions. In this way the number of possible alternative explanations can be reduced. This is often the approach taken in articles published in high-rated journals (e.g. Journal of Language and Memory, Journal of Experimental Psychology).

5.7 Summary

In this chapter we discussed the final step in designing and conducting an experiment: data analysis. We have shown that even in this stage of the experiment, the unintended variability in the data must be kept as small as possible. This is done, first, by carefully screening your data on extreme outliers, and regarding these data points as missing data. Second, data can be trimmed by calculating relative cutoff-values for each participant and replacing RTs larger (smaller) than this upper (lower) cutoff with this cutoff value. After the data have thus been screened and trimmed, participant and item means are computed for each condition. Since you not only wish to generalize over participants, but also over items, statistical analyses must be carried out on the participant as well as the item means. We have illustrated how to analyze RT data using an SPSS repeated measures analysis and how to interpret the SPSS output.

On the basis of the experimental results you can determine whether your prediction is borne out or not. Bear in mind that when the prediction is borne out, the underlying hypothesis is not necessarily true: often the results can be explained by various other hypotheses. The strongest conclusion one can draw on the basis of one experiment is that the results *support* the hypothesis, which is something different from saying that the hypothesis is true.

Chapter 6 Essential Steps in Carrying Out and Reporting an Experiment

In this text, we have been concerned with the steps which are necessary when you carry out an experiment. In this chapter, we will summarize these steps. These steps are in many ways parallel to the structure of a report of the experiment. In the second part of this chapter, we will give an outline of the typical format of a psycholinguistics article which reports the results of an experiment. You can use both of these as checklists while preparing, conducting and reporting your own studies.

Note: the check list used below is one appropriate for an experiment in which the participants see multiple conditions, and in which each item appears in multiple conditions across the experiment. It can easily be adjusted to different situations. For instance, if items only appear in one experimental condition, the analysis by items can be omitted.

6.1 The Experimental Procedure

6.1.1 Constructing Hypotheses (cf. Ch. 1)

1. Specify your research question.

2. Specify possible answers to this question (hypotheses).

3. Determine what *specific* predictions these hypotheses make in specific sorts of materials
   - What are your experimental conditions?
   - What differences do you predict?

4. Specify the factorial design of your experiment (cf. also Chapter 5).

6.1.2 Constructing Materials (Ch. 1-3)

1. Choose the most appropriate experimental technique to measure the precise differences you expect.

2. Think about which factors might cause artifacts and noise and how to avoid them.

3. Make a lot of potential experimental item sets (number of conditions x 15 or more). Check for repetitions; replace repeated content word if possible.

4. Pretest the items on plausibility, or other subjective aspects by having a panel of persons rate them on a scale. This is actually an experiment on its own. The steps you have to take are comparable to the ones described in this section. We will abbreviate them here. See below for more details.

- divide the item sets into as many item groups as you have conditions;
- construct filler (distractor) items
- distribute materials across lists in a Latin Square
- create blocks (Ch. 3)
- make and check the participant lists
- make instructions
- collect data
- determine mean and standard deviation of the rating per item and per condition
- select the best items for your 'real' experiment.

5. Consider whether there is any other factor that might confound your results. Pretest for possible effects of that factor, if necessary.

6. Obtain length and frequency information of the critical words (words being measured) using an electronic data base.

7. Match the materials in the various conditions on length, frequency, plausibility and any other factors that may be relevant, using Excel or some other spreadsheet. When you have found the perfect match, save the means and standard deviations (you will need them for your report). Even if the condition seem to be perfectly matched, test this statistically

8. Divide the sets into item groups, matched on length, frequency, plausibility, etc. This information about the matching (how tested, means, standard deviations) goes into your report as well.

-->> Note: make back-ups after each step! <<--------

6.1.3 Preparing Materials (Ch. 3, 4, Appendix 5)

1. Construct filler (distractor) items. Check for repetitions and replace where possible.

2. Consider how the items and conditions are assigned to participant lists according to a Latin square design. The next steps are followed to make a 'mother list' from which the participant lists are generated.

3. Code the experimental items, coding each item set and condition, e.g.:

Item#	condition	description	item
4	a	plaus	I take my coffee with cream and sugar.
4	b	implaus	I take my coffee with cream and dog.

4. Code the distractor items if necessary.

5. Write down which item set (coded how) is going to appear in which condition (coded how) in which participant list (coded how) according to your Latin square.

6 Make blocks of fillers and experimental items as described in Chapter 3. Note that this "mother list" still contains every member of each set. The order is not scrambled yet.

Block	Item#	condition	description	item
1	4	a	plaus	I take my coffee with cream and sugar.
1	4	b	implaus	I take my coffee with cream and dog.
1	5	a	plaus	I spread the warm bread with butter.
1	5	b	implaus	I spread the warm bread with socks.
etc				
2	8	a	plaus	She put the candy in her mouth.
2	8	b	implaus	She put the candy in her air.
etc				

7. Construct this mother list according to the format you need to insert the items into Eprime.

8. Check (and re-check) the mother list for typos

9. Derive your participant lists from the mother list, picking the correct member (condition) from each item set, according to your Latin square design. Put each participant list in a different worksheet. Name each sheet such that you can see which participant list it contains. For instance, 'ek_List1', 'ek_List2'. 'ek' is the name of the experiment.

10. Check and double check each participant list file on typos, whether the right condition has been selected and whether the coding is correct.

11. Make a block of practice item, code them as necessary and store them as a separate worksheet.

12. Copy the lists into E-prime. Make separate _es files for each list. Each of the lists contains the same practice block.

-->> NB: make back-ups after each step! <<--------

6.1.4 Preparing to Run the Experiment (Ch. 4, Appendix 5)

1. Check whether each file is run correctly by Eprime .

2. Check whether all and only those responses are stored which you are interested in.

3. Think about instructions and experimental procedure: print out a leaflet with instructions so you can let the participant read them.

4. Make a log book. Fill in which list (file) each participant is going to read. Make sure that there an equal number of participants are assigned to each list.

5. Run a practice participant (preferably a very critical person) and do an extensive debriefing. Watch very carefully how s/he is carrying out the task and whether s/he is following the instructions. Do the instructions need to be expanded/revised?

-->> NB: make back-ups after each step! <<--------

6.1.5 The Actual Experiment (Ch. 4, Appendix 5)

1. Run your 'real' participants
   - Obtain informed consent
   - Complete the log sheet; obtain data related to sex, gender, language etc.
   - Instruct the participant
     - have the participant read the written instructions
     - go over most important details of the instructions with the participant before starting the practice block
   - double check the list number and participant number before starting E-prime
   - Run the practice; observe whether participant is carrying out the task properly
   - Leave the participant alone while s/he is doing the actual experiment
   - After each participant:
     - debrief your participant and note any interesting observations
     - make a back-up of the _.edat file
     - complete the log sheet

-->> NB: make back-ups after each step! <<--------

6.1.6 Data Analysis (Ch. 5, Appendix 5)

1. Concatenate your _.edat files using E-Merge

2. Using E-DataAid, Select the columns you need; export to SPSS format

3. Inspect the data for outliers; drop participants if necessary

4. Perform absolute and/or relative cutoffs

5. Create item means and participant means (in SPSS)

6. Conduct ANOVA, Repated Measures or other tests (SPSS) on both item and participant means.

7. Read the output to see which effects are significant. Examine the magnitude and direction (sign) of the differences found, inspecting the means per condition.

8. Are your predictions borne out?

-->> NB: make back-ups after each step! <<--------

6.2 How to Report Your Experiment

This is a common outline of what a report of an experiment should look like.

- The goal of a report of an experiment is to make it possible for the reader to carry out essentially the same experiment in their own laboratory using their own equipment as a ***replication*** (without ANY potentially significant differences in materials of presentation) in order to see if the same result is found:

THINK what information you would need to do that.

It is also necessary to inform the reader about any detail that indicates to what extent the results are to be trusted.

For example, if you did not match the materials on certain factors, the results might not be reasonable, so explain what was matched and how the match was carried out. When in doubt, include!

Last of all, it should be absolutely clear
- why you predicted a particular result, and
- what conclusions you draw from the results.

It is important that you report all of the points that are made here, but there is nothing sacred about this order. You should think carefully about how you can present your experiment most clearly. Another organization is better if it makes the experiment and results clearer.

6.2.1 Introduction

- Specify your research question (give references to existing literature).
- Specify possible answers to this question (hypotheses)
- ***Argue*** for the specific predictions these hypotheses make about the processing of specific types of items
- ***Show how*** the experimental technique measures the precise differences you expect.

6.2.2 Method

6.2.2.1 Materials

- Give examples of your experimental items and conditions. Describe how many items of each sort you have tested.
- Describe the factors on which you have matched the conditions and how you obtained the data on the basis of which the conditions were matched.
- Provide **data (means and standard deviations)** concerning the matching between conditions.
- Describe how you assigned items to participant lists (Latin Square), including matching of item groups (means and standard deviations).

6.2.2.2 6. Participants

- Describe your participant sample.

6.2.2.3 6.Procedure

- Describe the task:
  - what does the participant see/hear?
  - How fast is the presentation
  - how and when does s/he have to react?
  - what happens step by step?
- Describe the instructions .

6.2.3 Results

- Specify how data has been treated (absolute cutoff, outlier replacement), that you have carried out participant and item analyses, which statistical tests you used and why you used these tests .
- Specify the factors in your design (cf. also Chapter 5) for each analysis (e.g. participant and item analysis) .
- Which effects are significant? Provide df, F- and p-values and a table/figure containing the means per condition .

6.2.4 Discussion

- Summarize the argument from steps 1-3.
- Are your predictions borne out?.
- Discuss any other hypothesis that can potentially explain the results.
- What more do you need to know to really answer your research question? Suggest further experiments.

6.2.5 Appendices

- Include your experimental materials.
- Provide data concerning matching between item groups if that has not been included in the body of the report.

References

Brouw, J. (1995) 'Animacy in Sentence Processing.' Doctoraalscriptie. RUG, Groningen.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition, 10,* 65-70.

Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior, 8,* 457-462.

Koster, Ch. (1993) *Errors in Anaphora Acquisition.* Doct. Diss. OTS, Utrecht.

Kutas, M. & S.A. Hillyard (1980) 'Reading Senseless Sentences: Brain Potentials Reflect Semantic Incongruity.' In: *Science, 207, 203-206.*

Marslen-Wilson, W. (1985) 'Speech Shadowing and Speech Comprehension.' *Speech Communication, 4,* 55-73.

Marslen-Wilson, W. and L.K. Tyler (1980) 'The Temporal Structure of Spoken Language Understanding.' *Cognition, 17,* 209-242.

Marslen-Wilson, W. and A.Wlesh (1978) 'Processing interactions and lexical access during word recognition in continuous speech.' *Cognitive Psychology, 10,* 29-63.

Mehler, J., J. Segui and P. Carey (1978) 'Tails of Words: Monitoring Ambiguity.' *Journal of Verbal Learning and Verbal Behavior, 17,* 29-35.

Neville, H, J. Nicol, A. Barss, K.I. Forster & M.I. Garrett (1991) 'Syntactically Based Sentence Processing Classes: Evidence from Event Related Brain Potentials.' *Journal of Cognitive Neuroscience, 3,* 151-165.

Stowe, L.A., A.A. Wijers, A. Willemsen, E. Reuland, A.M.J. Paans and W. Vaalburg (1994) 'PET Studies of Language: An Assessment of the Reliability of the Technique.' *Journal of Psycholinguistic Research,* 23, 499-527.

Swinney, D.A. (1979) 'Lexical Access During Sentence Comprehension: (Re)Consideration of Context Effects.' *Journal of Verbal Learning and Verbal Behavior, 18,* 645-659.

Tanenhaus, M.K., J.L. Leiman and M.S. Seidenberg (1979) 'Evidence for Multiple Stages in the Processing of Ambiguous Words in Synatctic Contexts.' *Journal of Verbal Learning and Verbal Behavior, 18,* 427-441.

Trueswell, J.C. , M.K. Tanenhaus and S.M. Garnsey (1993) 'Semantic Influences on Parsing: Use of Thematic Role Information in Syntactic Ambiguity Resolution'. *J. of Memory and Language, 33,* 285-328.