Later published as:
Myers, J. (2009). The design and analysis of small-scale syntactic judgment experiments.
Lingua, 119,425-444. doi:10.1016/j.lingua.2008.09.003

The design and analysis of small-scale syntactic judgment experiments

James Myers
National Chung Cheng University
Lngmyers@ccu.edu.tw

Resubmitted March 3, 2008

Abstract

A growing literature argues that native-speaker syntactic judgments are best tested with standard psycholinguistic protocols, but the difficulty of formal experimentation has prevented many syntacticians from trying it. Fortunately, as shown by a textual analysis of a typical theoretical syntax paper (Li 1998), syntacticians already recognize the value of proper experimental designs and quantitative analyses, albeit in nascent forms. Small-scale judgment experiments build on this foundation by applying the minimum amount of extra effort needed to draw valid statistical inferences from the type of data most familiar to syntacticians, namely binary judgments involving very few of speakers and sentences. Statistical methods appropriate for such data, some hitherto underused, are described, and a software tool for automating small-scale judgment experimentation is introduced. The simplicity and power of the methods are then illustrated in a small-scale test of the claims in Li (1998) on naive Chinese speakers.

Keywords: Experimental syntax; Methodology; Quantitative linguistics; Chinese

1. Introduction

From the very beginnings of generative syntax, there have been calls to supplement, or even replace, informally collected native-speaker judgments of well-formedness with formal experiments of the sort used in the rest of the cognitive sciences (e.g. Hill 1961). Only within the last decade, however, have formal judgment experiments gone truly mainstream. Sparked by the publication of Schütze (1996), Bard, Robertson, and Sorace (1996), and Cowart (1997), there has been an explosion of studies testing grammatical hypotheses against judgments collected according to the standard protocols of experimental psycholinguistics. These protocols include the use of naive speakers to avoid experimenter bias, large sample sizes to improve representativeness, randomized presentation order to avoid confounds due to fatigue or cross-item priming, counterbalanced lists and filler items to prevent speakers from making explicit comparisons across items, ordinal or continuous-valued judgment scales for greater sensitivity, and standard statistical tests like analysis of variance (ANOVA). Recent examples of such studies published in prominent linguistics journals include Bernstein, Cowart, and McDaniel (1999), Featherston (2005ab), Clifton, Fanselow, and Frazier (2006), and Sprouse (2007).

Despite their apparent advantages, formal experiments have not been adopted as general practice by theoretical syntacticians. One reason for this is that such experiments take a greater amount of time and effort than really seems necessary to test many judgment claims. As Labov (1975:81) notes, "[i]f every linguistic fact had to be examined by representative sampling, experiment and observation, we would never proceed beyond the simplest patterns of the most well known languages." Similarly, Phillips and Lasnik (2003:61) write that the "[g]athering of native-speaker judgments is a trivially simple kind of experiment, one that makes it possible to obtain large numbers of highly robust empirical results in a short period of time, from a vast array of languages." Even experienced experimental syntacticians fall back on informal judgments on occasion: Schütze (2001) cites only informal judgments, and Bernstein et al. (1999) supplement their formal judgment experiment with further informal judgments of their own. The implication seems to be that formal experimentation should be reserved only for particularly subtle patterns that informal methods cannot resolve.

Yet if only formal experimentation is powerful enough to tell for sure whether a claim is valid, how can we trust informal intuitions to tell us when formal experimentation is necessary? The solution to the paradox, I argue, is to take seriously the assumption, made both by advocates like Phillips and Lasnik (2003) and by critics like Labov (1975), that informal judgment collection is itself a form of experimentation. Since informal methods and full-fledged experimentation lie on a continuum, rather than representing radically different types of data sources, there must be a form of experimentation somewhere in between, one that is almost as simple and quick as informal methods, but which relies on the careful design and quantitative analysis of formal experimentation. Because such experiments could be run quickly and easily, syntacticians wouldn't have to restrict their use to especially difficult cases, but could use them any time they (or their critics) have doubts about empirical claims. I call this type of experiment a small-scale judgment experiment.

I motivate and illustrate this concept as follows. Section 2 shows, through a systematic analysis of the example sentences in a representative syntax paper, Li (1998), that theoretical syntacticians already understand the need for proper experimental designs and quantitative argumentation, even if they apply them only informally and inconsistently. Section 3 builds on this foundation, providing a brief tutorial on the design of small-scale experiments involving binary yes/no judgments from very small samples of speakers and sentences, as well as the statistical tests (many underused) that are most appropriate for them. Since even simple procedures can be made still simpler through automation, I then describe MiniJudge, a free, online, open-source program for designing, running, and analyzing small-scale judgment experiments (Myers 2008a). Finally, I illustrate the statistical methods on real data from a small-scale experiment testing the judgments in Li (1998) on naive Chinese speakers.

2. The design of informal syntax judgment experiments

The discussion in this paper starts from two key assumptions, namely that the traditional collection of syntactic acceptability judgments represents a genuine, if informal, method of psycholinguistic experimentation, and that judgment experiments run in accordance with standard psycholinguistic protocols are inherently more sensitive as well as inherently more convincing, especially to potential research collaborators in the other cognitive sciences (for defenses of one or both of these assumptions, see for example Adli 2005, Clifton et al. 2006, Cowart 1997, Featherston 2005a, 2007, Phillips and Lasnik 2003, Schütze 1996, and Sproat 2007).

While these two points have been advocated for many years now, discussions of how to improve the quality of judgment data have focused on details of implementation, such the choice of speakers (e.g. Labov 1975), instructions (e.g. Schütze 2005), and measurement scale (e.g. Bard et al. 1996). The discussion in this paper focuses instead on the core properties defining true experiments as understood in psycholinguistics and most other sciences, namely their use of logical designs and quantitative analyses. I demonstrate that even informal judgment collection has these characteristics, albeit in a nascent form. These then provide the seeds for the more rigorous, yet still relatively simple, methods of small-scale judgment experimentation.

In this section, I first review the logical principles of experimental design and analysis (2.1), and then examine a representative theoretical syntax paper, Li (1998), to see how these principles are and are not applied (2.2).

2.1 Experimental design

Psycholinguistic experiments are characterized by controls, factorial designs, and quantitative analyses. These features are not accidental, but essential to their power.

In ordinary language, an "experiment" is any attempt to try something and see what happens, but in most sciences today the term has a narrower meaning. The essential element of an experiment is a comparison of the output effects associated with different types of inputs. This is why control conditions are essential to experimental design. This holds of judgment experiments as much as any other type, since an acceptability judgment is a type of psychological "sensation" (Schütze 1996:52) that falls along a continuum (Chomsky 1965, Bard et al. 1996), without any built-in reference points. Thus just as the minimal pair is the phonologist's response to the problem of phonetic gradience, syntacticians generally recognize that minimal pairs are necessary in order to test whether some structural difference in otherwise identical sentences correlates with a difference in acceptability.

However, in most sciences, including psycholinguistics, researchers prefer to look not at minimal pairs, but minimal sets, defined not by just one factor, but by two or more factors crossed so that all possible combinations can be tested. Factorial designs were first explicitly advocated over eighty years ago by the statistician G. A. Fisher in a classic (and nontechnical) paper (Fisher 1926), and they are now ubiquitous. Multi-factor experiments are not only more efficient than single-factor experiments, since several questions can be asked at once, but they make it possible to understand how the factors interact with each other. Interactions are common not only in psycholinguistic processing, but the structure of competence as well, since linguistic claims often concern the relationship between elements, where each element is defined by a separate factor. A well-studied example is the *that*-trace effect in English (Chomsky and Lasnik 1977), which involves not only the presence vs. absence of *that* (one factor) but also whether the extraction site is in subject or object position (another factor); the *that*-trace itself is reflected in the interaction (subject extraction is disfavored in the presence of *that*). Cowart (1997) provides extensive discussion of formal judgment experiments of the *that*-trace effect that crucially rely on this factorial design.

A final feature of psycholinguistic experiment is statistical analysis. As Fisher also emphasized, a statistical analysis is not a rhetorical flourish tacked onto the end of an experiment, but rather it is essential to the whole logic. Without a proper design, a statistical analysis isn't very useful. For example, if ten people judge an isolated sentence (without a comparison control) and all of them find it acceptable, this is a statistically significant result (as we will see in section 3.1), but what does it mean? It may be that the sentence actually is acceptable, but it could mean instead that the judges have a positive response bias (as do many speakers; Crain and Thornton 1998:213).

Equally important is the complementary point, that experiments need to be designed with statistics in mind. Fisher (1926) cites the example of an experiment in which we fertilize one acre of ground and leave another untreated, and find that the first has a greater crop yield. This is only convincing evidence of the efficacy of the fertilizer if we can be sure that the two acres are otherwise identical, but in the real world this is rarely the case. A similar situation holds in syntax, where as Cowart (1997:47) puts it, "any one person's response to any one sentence is usually massively confounded": we can't tell if the response was due to the theoretically interesting syntactic properties, or to idiosyncratic properties of the speaker or sentence. The solution proposed by Fisher (and now accepted as standard practice in virtually all sciences) is to test multiple exemplars of each factor combination, assuming that they vary randomly within each such cell of the design. Only if the effects of the theoretically interesting factors stand out over this random noise do we have any hope of convincing a skeptical critic of our claims.

Psycholinguists may apply these three aspects of experiments with a high degree of sophistication, but as will be shown next, syntacticians are also aware of their importance.

2.2 A case study: Li (1998)

Phillips and Lasnik (2003:61) are partially right to emphasize that "appropriate control items" are part of "[a]ny good linguistics study," and factorial designs are not uncommon as well. Syntacticians even use an implicit sort of quantitative analysis. Neither the experimental designs nor the statistics are applied with the rigor required in the psycholinguistic literature, but the fact that they are there at all means that they can be built on with a minimum of extra effort.

I illustrate these points with a close textual analysis of a single syntax paper, Li (1998). This paper is chosen for three primary reasons. First, it is short, which makes a thorough analysis feasible. Second, it is an example of a "good linguistics study." Not only does it have far more citations on Google Scholar (scholar.google.com, consulted on February 21, 2008) than any of the other nine syntax squibs published in the same journal in the same year (44; the runner-up has only 17), but the author has a reputation for being methodologically careful, and has lectured on syntactic methodology at universities across the world (e.g. Li 2004).

Third, and most important, the paper's methodology is entirely typical of the theoretical syntax literature. It relies solely on acceptability judgments (presumably generated by the author herself) measured on a binary good/bad scale, in order to support a large number judgment claims, many of which are quite subtle, in a language, Chinese, that is likely to be unfamiliar to most readers.

The binary judgment scale remains overwhelmingly preferred in the syntax literature, despite recent arguments for finer-grained scales (e.g. Bard et al. 1996, Cowart 1997, Featherston 2005ab, Sorace and Keller 2005, Sprouse 2007; though see Weskott and Fanselow 2008 for counterarguments). For example, of the 1812 distinct sentences or phrases cited in the 32 syntax and semantics papers published in Volume 37 of *Linguistic Inquiry* (2006), only 96 (5%) are marked with a judgment diacritic other than blank (accept) or * (reject), and even when other judgment diacritics are used, the focus remains on binary contrasts (e.g. ? vs. ???, or blank vs. *?, as in Ishii 2006:158).

Chinese is not only unfamiliar to most readers, making it impossible for them to confirm Li's judgments, but it is a language in which judgments have been notoriously controversial. For example, many of the judgments in Huang (1982) have been challenged by other native speakers, including speakers from the same dialect region (e.g. Chen and Pan 2003, Lee 1986, Shi 1994, Tang 1984, Xu 1990, 1996), and the same is true for some of the Chinese judgments in Aoun and Li (2003) (e.g. Ou 2006 rejects Aoun and Li's example (2b), p. 133).

Finally, the empirical claims in Li (1998) are both subtle (relating to non-obvious semantic contrasts) and relatively numerous compared with the typical psycholinguistics paper; depending on how one counts, there is roughly one major factual claim per page. Put together, these features make Li (1998) a useful test case: it is both typical of the "best" of the syntactic literature, yet simultaneously represents what critics see as the "worst" aspects of traditional syntactic methodology.

The central theoretical claim of Li (1998) is that number expressions in (Mandarin) Chinese can have both quantity-denoting and individual-denoting interpretations, a difference that has a variety of empirical effects (Li unifies these effects via certain theoretical constructs that are irrelevant here, since my focus is on judgment claims, not the higher-level claims putatively derived from them). Evidence for these effects come from binary acceptability judgments (i.e. informal judgment experiments) on thirty-eight Chinese sentences, including variants assuming a specified interpretation or discourse context (see examples below).

Li places these sentences in five distinct types of implicit experimental designs. The design most like that found in psycholinguistics is the factorial design involving two binary factors; Li applies this in support of two claims (with data from eight sentences or sentence variants). The first claim is that quantity-denoting expressions cannot bind reflexives. This is supported by a claimed interaction between two binary factors, one indicating whether the noun phrase closest to the reflexive is a number phrase, as in (1a) vs. (1b), and the other indicating whether the reflexive should be interpreted as coreferential with this noun phrase or the earlier one (Chinese permits long-distance reflexives), as indicated by the subscripts on the reflexive *ziji*.[1] No claim is made about each factor separately; the crucial point is that the

[1] The original example numbers from Li (1998) are given in square brackets; see also Table 1. Li's morpheme glosses are eliminated because the experimental designs are clear without them (Chinese and English have similar word orders). The phrases defining the factorial contrasts are underlined in the Chinese and English versions. The acceptability diacritics (blank vs. *) are Li's.

two factors interact with each other.

(1) a. Zhangsan$_i$ zhidao sange ren$_j$ yiding ban-de-dong ziji$_{i/*j}$ de gangqin.
"Zhangsan knows that three people certainly can move self's piano." [=(22a)]

b. Zhangsan$_i$ zhidao Lisi$_j$ yiding ban-de-dong ziji$_{i/j}$ de gangqin.
"Zhangsan knows that Lisi certainly can move self's piano." [=(22b)]

The second claimed interaction also involves two binary factors. One represents the presence/absence of the existential marker *you* "have, exist", claimed to force an individual-denoting interpretation on number expressions. The other factor represents the scope of a higher number phrase over a lower one, giving rise to an interpretation where the two numbers are multiplied (see Li's (9) vs. (23a)). Again the crucial theoretical claim concerns the interaction: only with the presence of *you* can the higher number phrase have scope over the lower one.

A more common design in Li (1998), however, involves minimal pairs; four judgment claims are supported with this design (sixteen sentences or sentence variants). Two examples will suffice to show why such designs tend to be avoided in psycholinguistics. Li starts the paper by reviewing the familiar (to Chinese syntacticians) complementary distribution shown in (2), where a number expression is disfavored in subject position in (2a) but saved by adding *you* in (2b). Li later argues that this is because *you* forces an individual-denoting interpretation in (2b).

(2) a. *Sange xuesheng zai xuexiao shoushang le.
"Three students were hurt at school." [=(1)]

b. You sange xuesheng zai xuexiao shoushang le.
"There are three students hurt at school." [=(3)]

A second minimal-pair claim is merely hinted at in a footnote (footnote 3, p. 694), where Li agrees with a Chinese-speaking colleague that sentences like (2a) improve "if they are answers to *how many* questions." This implies a binary factor representing null context vs. a context forcing a quantity-denoting interpretation.

The problem is that these two minimal pairs partially overlap. They both contain (2a), but do not complete the paradigm by also considering (2b) in both null vs. *how many* contexts. In other words, an incomplete two-factor experiment is implied, as illustrated in (3).

(3)	Factor 1	Factor 2	Example	Judgment
a.	[-you]	[-how many]	(2a) in text	Bad
b.	[-you]	[+how many]	(2a) in footnote	Good
c.	[+you]	[-how many]	(2b) in text	Good
d.	[+you]	[+how many]	(not available)	Unknown

This quasi-design poses problems for the interpretation, since it's no longer clear what the control condition is. For example, is it meaningful to compare the judgments for rows (3b) and (3c)? An incomplete factorial design also cannot be analyzed statistically, since the two pairs in (3a) vs. (3b) and in (3a) vs. (3c) are not independent (i.e. the variability in potential responses won't be partitioned in a mutually exclusive way). Moreover, interactions are impossible to detect in an incomplete factorial design. This may not only cause important insights to be missed, but if an interaction exists and isn't taken into account, it may be that what looks like a "main" effect is actually due to this hidden interaction.

The twenty-four sentences or sentence variants used in the factorial and one-factor designs described above represent the majority (63%) of the examples cited in Li (1998); these designs are essentially identical to designs commonly found in the psycholinguistics literature. Unfortunately, this does not hold of the remainder of Li's example sentences. Thus in the third type of implicit design, a two-factor experiment is cobbled together from two single-factor experiments, using unmatched sentence pairs. There are at least three instances of this design logic in Li's paper (it's impossible to count them precisely, since Li doesn't always make the intended contrasts explicit). One of them involves the comparison she highlights between sentences like (2) with those in (4), where adding *you* "have" has precisely the opposite effect on acceptability. She ascribes this contrast to the quantity term *gou* "enough" in (4), which itself makes the number expression quantity-denoting and so competes with *you*.

(4) a. Sanzhi gunzi gou ni da ta ma?
"Are three sticks enough for you to him (with)? [=(8)]

b. *<u>You</u> sanzhi gunzi gou ni da ta ma? [=(17a)]
[no gloss given]

The quasi-factorial design implied by putting (2) together with (4) is actually even more problematic than the design in (3). The design in (3) could be fixed simply by filling the empty cell in (3d) with a properly matched sentence, but if two distinct pairs are used, like (2) vs. (4), one of the theoretically interesting factors (here, the absence vs. presence of *gou* "enough") is inextricably confounded with all of the other uninteresting properties that make the pairs different (e.g. statement vs. question). A more convincing way to argue that *gou* is the crucial element would be to use a true factorial design, with quadruples of matched sentences, as Li does in (1) for reflexive binding.

Despite some problems, the designs in Li (1998) described so far are (mostly) consistent with the assertion in Phillips and Lasnik (2003) that all "good linguistics studies" use "appropriate control items." In fact, however, Li (1998) is typical of the syntax literature in also citing judgments for isolated examples, without explicitly mentioning any controls for comparison. One of several examples of this is shown in (5), introduced in Li's paper as a counterexample to the pattern seen in (2).

(5) Liangzhang chuang (, wo tingshuo,) ji le wuge ren. Na shizai shi tai ji le.
"Two beds (, I heard,) were crowded with five people. That was really too squishy."
[=(5)]

It is argued that the claimed acceptability of (5) follows from the use of the number expression as quantity-denoting, since the first sentence "concerns quantity, rather than (the existence of) some individuals" (Li 1998:695). However, unlike the *gou* sentences in (4), no control sentences are ever cited, so it is impossible for non-native speakers to be sure that its acceptability would drop if *you* were added, as Li's claim would predict. Moreover, there are hints that the status of (5) as acceptable is not entirely uncontroversial, given the use of contextual clues (the follow-up sentence and the optional parenthetical).

In defense of Li (1998) (and the syntax literature generally), such uncontrolled examples do not seem to be in the majority. It may also be possible to argue that even isolated examples have implicit controls, namely the corpus of "good" and "bad" sentences encountered by a native speaker over the course of a lifetime (which, of course, is not available to non-native speakers). Whitman (2002:86-95) formalizes this logic in a full-fledged judgment experiment, arguing that sentences of a certain type are grammatical because a large proportion of speakers gave judgment scores for them falling within the same range they used for uncontroversially grammatical filler sentences. One serious weakness of this logic, addressed again below, is that it depends not on a significant difference, but on the failure to find a difference, a failure that may actually arise from insensitivity of the test.

The final type of experimental design used in Li (1998) is the hardest to justify. The above designs all rely on what Wasow and Arnold (2005) call primary intuitions, where a speaker evaluates the acceptability or meaning of a sentence. However, Li also makes use of what they call secondary intuitions, in which speaker-linguists evaluate why a sentence has a given acceptability level or meaning. Secondary intuitions are clearly problematic as data, since they confound the roles of experimenter and subject and rely on the thoroughly falsified notion that introspection into internal mental processes is reliable (Nisbett and Wilson 1977).

Li seems to be using a secondary intuition in her explanation for the acceptability of (5) quoted above, since it relies on her otherwise unjustified analysis that it "concerns quantity". Another example relates to Li's claim that quantity-denoting expressions cannot be coreferential with a sentence-external discourse element. Li argues that an apparently acceptable instance of this structure "need not provide a counterexample because the pronoun can refer independently to a group of people that happens to consist of three members" (footnote 10, p. 699). It is not obvious (especially to non-native speakers of Chinese) how to interpret this suggestion in a way that isn't question-begging. It seems to stipulate that the number phrase in the acceptable counterexample is individual-denoting (hence can be coreferential), but why this interpretation is possible here, when the same number phrase is unacceptable in a sentence with apparently the same structure and discourse context, is precisely what is at issue.

Again, apparent cases of secondary intuitions seem to be relatively rare in the syntactic literature (though a systematic survey has yet to be conducted). Moreover, once the problem is recognized, the solution to it is relatively straightforward: test (or at least imaging testing) one's examples on naive native speakers. If the researcher cannot think of a way to make the necessary distinction clear to nonlinguists by modifying the wording or placing the sentence in the proper discourse context, then the distinction might relate to a secondary intuition and should be reconsidered.

This systematic examination of data use in a single paper confirms that syntacticians do indeed appreciate the principles of proper experimental design, even if they do not always apply them consistently. As has already been noted in the literature, linguists should be careful to avoid hidden secondary intuitions, and should reduce the use of uncontrolled isolated examples (unless they are conducting corpus analyses, which have their own principles of design and analysis; see e.g. Manning and Schütze 1999). A less often discussed problem (though see Cowart 1997) is the overuse of minimal pairs. Given how how often linguistic factors interact, linguists should learn to think in terms of complete, well-structured factorial designs.

So much for experimental design. What about quantitative analysis? Here there is much less to say, since like most theoretical syntax papers, Li (1998) makes no explicitly quantitative statements. The examples cited in the paper play double duty as representatives of whole classes of sentences, like the examples that psycholinguists cite in the text of their papers, but presumably they also constitute the author's primary data. Nevertheless, it is likely that other examples were considered but were not cited due to lack of space. If the selection process involved sifting sentences to find "clear cases" (a strategy endorsed, among other places, in Chomsky 1957:14 and Labov 1975:103), we may have to worry about confirmation bias (Nickerson 1998), whereby hypotheses are tested by looking for supporting examples rather than counterexamples. Fortunately, sampling in a less biased fashion does not require much extra effort; Cowart (1997:50-51) suggests using a thesaurus to come up with lists of semantically related verbs, and generating sentences around them.

A more overt sign that quantitative argumentation is relevant in Li (1998) is the fact that Li cites five different counterexamples to the familiar *you* generalization in (2). A single counterexample might be written off as a statistical fluke, but five begin to look like a pattern (which she then analyzes). Another bit of implicitly quantitative logic is that minimal pairs are never used alone. Each of the four minimal pairs cited in the paper is itself given a "redundant" pair, where a second doublet of sentences illustrates precisely the same contrast.

Unfortunately, when quantitative arguments are made more explicitly in syntax papers, they are not always conducted effectively. For example, in Soh (2005), another Chinese syntax squib published in the same journal, it is reported that of eleven speakers consulted on a Chinese sentence (her (25), p. 151), six accepted it while five did not (p. 151, fn 9); the sentence is thus marked "(?)". Yet not only does Soh continue to assert the grammaticality of this sentence without any control (closely similar but less acceptable sentences are quickly dismissed as theoretically irrelevant), but the quantitative results do not even support her claim: the observed six-vs-five split is consistent with random judgments, as if the speakers were flipping a coin.

In short, syntacticians do respect experimental design and quantitative analysis, even if they may not apply them consistently or effectively. The gap between informal methods and full-fledged experimentation is thus not as great as often assumed. The rest of this paper shows how the gap may begin to be bridged.

3. The statistical analysis of small-scale judgment experiments

Theoretical syntacticians want to test whether their judgments of a sentence set generalize across other speakers and sentence sets, but they virtually never use statistics to do so. This is partly because of the traditional fear of mathematics in the humanities, but partly because statistics textbooks tend to give the misleading impression that statistics is only for continuous measures, like reaction time or voice onset time. Woods, Fletcher, and Hughes (1986:1) even go so far as to list the fact that syntactic judgments are "either-or decisions" as a reason why there is "no place here for statistics."

In fact, statistical techniques for analyzing binary yes/no judgments not only exist, but are perhaps the most conceptually simple of all statistical tests. In this section I describe techniques for one-factor experiments, techniques for two-factor experiments, and finally a technique that is the most powerful and general method for analyzing small-scale judgment experiments currently available. Some of these techniques will be familiar to the statistically sophisticated reader, but perhaps not all of them. All of the statistical analyses work well with very small samples, and all can be run with the free R program (R Development Core Team 2008), which is becoming a standard among quantitative linguists (e.g. Baayen 2008, Johnson

2008).

3.1 One-factor experiments

The simplest judgment experiment capable of providing statistical evidence for a syntactic claim (as opposed to an overall bias to say "yes" or "no") would have a single binary factor, contrasting a single experimental sentence with its matched control sentence. Suppose we test this sentence pair on multiple speakers. How do we determine if the pattern we get is too unlikely to have occurred by chance alone? More precisely, the question is how to calculate the *p* value, which represents the probability of getting a result at least as extreme as that actually observed (statistical significance is generally defined as $p < .05$, following an arbitrary convention introduced by Fisher).

Before answering this question, some basic issues must be highlighted for readers unfamiliar with statistical argumentation (further discussion may be found in any introductory statistics textbook). First and most fundamentally, it is not wise to fetishize *p* to the exclusion of all other values. In particular, "statistically significant" does not necessarily mean "significant" in the ordinary sense, since the effect may be very small (see Cowart 1997:123 for a syntactic example). Second, finding $p > .05$ does not mean that a result can be dismissed as chance. Linguists should be particularly qualified to see why, since the logic hinges on the scope of negation. Namely, $p > .05$ only means that we did not find a pattern (at the conventional significance level); it does not mean that we found that there is no pattern. Similarly, finding $p < .05$ indicates that a result is unlikely to have happened by chance, but it doesn't rule out this possibility; in fact, by definition, $p < .05$ should be observed about 5% of the time through chance alone. Nevertheless, by the rules of the game as it is played in most sciences, a significant result tends to shift the onus on the critics to justify any continued skepticism. Finally, researchers generally prefer nondirectional *p* values (for technical reasons, more commonly called two-tailed *p* values), which measure the chance probability of getting an effect as least as big as what was observed, in both directions (e.g. both A > B and A < B, even if the research hypothesis predicts only A > B). Nondirectional *p* values provide a stricter test of the research hypothesis and thus are even more convincing to skeptics (see e.g. Kirk 1995:58).

To return to our question, the computation of *p* values in binary judgment experiments depends on how we distribute the sentences. The easiest method, practically speaking, would be to give both sentences to each speaker, in what is often called a within-groups design (here grouped by speaker). For any given speaker, there are four possible outcomes: YN, NY, YY, NN (where YN represents a "yes" judgment for one sentence and a "no" judgment for the other, and so on). Conveniently, McNemar (1947) showed that we only need pay attention to the discordant pairs (YN, NY), since the concordant pairs (YY, NN) can't provide positive information about a hypothesized difference. This means that if discordant pairs are generated by chance alone, there should be an equal number of YN and NY response pairs. The chance hypothesis is thus as if speakers were flipping a coin to choose between YN and NY response pairs; the greater the proportion of one type versus the other, taking the total number of discordant pairs into account, the less likely the results are due to chance. This makes the *p* value mathematically straightforward to calculate. The statistical test based on this insight is the exact McNemar test (also called the sign test), and it is easily computed (see Appendix A2).

Suppose, for example, that six Chinese speakers reject the sentence in (2a) with the number phrase without existential *you* but accept the matched version with *you* in (2b), and no speaker has the reverse judgment. This turns out to be significant by the exact McNemar test (nondirectional $p < .05$). In fact, this imaginary situation represents the smallest possible within-groups one-factor experiment (only six data points!) capable of giving a significant result.

However, since within-groups designs present contrasting items to the same speakers, they risk inducing task-specific strategies. For example, speakers may feel obliged to give complementary judgments merely because there doesn't seem to be any point to run the experiment otherwise (see Schütze 1996:79-80 for related problems). Researchers concerned about this may choose to hide the contrast by mixing irrelevant filler sentences into the judgment survey. Another option, however, would be to show each sentence in the pair to a different group of speakers. In this case, each judgment is independent of all the others, so now the coin flips represent individual judgments rather than paired judgments. An experiment of this type is thus called between-groups. Suppose, for example, that we collect judgments from seven Chinese speakers, three judging (2a) and four judging (2b), and all of the former reject (2a) while all of the latter accept (2b). Is this pattern statistically significant?

The relevant statistical technique for this situation is called Fisher's exact test (described by its inventor, in a famously informal manner, in Fisher 1935). What Fisher's test does, in essence, is compute the chance probability of getting a contrast between yeses and nos at least as big as that observed. More technically, it counts how many ways there are to get a contrast at least this big, and then divides this by the number of ways there are to distribute the observed data (in both cases, the number of sentences, yeses, or nos are kept constant, e.g. by exchanging a yes and no for sentences of one type while simultaneously exchanging a no and yes, in the opposite direction, for sentences of the opposite type). Calculating nondirectional *p* values in Fisher's test is easy to do (see Appendix A3). For example, the judgment pattern described above gives a nondirectional *p* value below .05. Again, this is the smallest experiment with this design capable of showing a significant result.

Note that the between-groups design requires a bare minimum of seven speakers, whereas the within-groups design requires a bare minimum of six. A between-groups design is recommended only if there is no other choice, since it does a worse job than the within-groups design at handling cross-speaker variability, and thus it is less sensitive. To take a more realistic example, Schmitz and Schröder (2002) chose this design because they were comparing semantic judgments for two matched passages, and they didn't want to ask the same speakers to judge both passages, for fear that judgments for one passage would influence judgments for the other. They ended up testing 47 speakers (23 for one passage, 24 for the other), and still didn't quite obtain statistical significance by Fisher's test (responses were 14 vs. 9 for one passage and 8 vs. 16 for the other, nondirectional *p* = .08). Yet if their results had come from a within-groups experiment, under the best-case scenario (with the maximum number of concordant pairs) there would be six (= 14-8) or seven (=16-9) discordant pairs, all going the same way, enough for a significant *p* value by the exact McNemar test (*p* = .03 or *p* = .02).

The techniques described here take very little practice to get used to, and as the examples above have shown, make it possible to get statistically significant results with very small samples.

3.2 Two-factor experiments

As we saw in section 2, factorial designs have important advantages over one-factor (minimal-pair) designs. Fortunately, it is possible to extend the exact McNemar test and Fisher's exact test to two-factor experiments, though these extensions may be unfamiliar even to many quantitative linguists.

To generalize Fisher's exact test for a two-factor between-groups experiment (separate speaker groups for each sentence in a quadruple), the first step is to code the two factors (F and G) and the interaction (FG) so that each is independent of (orthogonal to) the other two parameters. Thus [+F] = 1, [-F] = -1, and likewise for [G], and the interaction is represented as the product of [F] and [G], so [+F+G] = [-F-G] = 1 and [+F-G] = [-F+G] = -1 (see e.g. Kirk 1995:229-235). Using this coding, it is possible to compute separate *p* values for F, G, and FG in a small-scale judgment experiment using a technique called exact logistic regression (Agresti 2002:251-7). Regression analysis attempts to find an equation that best describes the relationship between observations and predictive factors, treating each factor in the context of the others so that their independent contributions can be distinguished. Though exact logistic regression is available only in a very few statistics packages (e.g. SAS; Derr 2000), Myers, Huang, and Tsay (2007) show how the same *p* values can be computed much more simply by building on Fisher's test (see Appendix A4).

For example, suppose that we test the four sentences implied by (1), spelled out separately in (8), giving each sentence to four speakers (for a total of sixteen speakers). The numbers of "yes" judgments in this hypothetical experiment are shown in (9), along with the factors defining the sentences.

(8) a. Zhangsan zhidao sange ren yiding ban-de-dong ziji de gangqin. [ziji = sange ren]
b. Zhangsan zhidao sange ren yiding ban-de-dong ziji de gangqin. [ziji = Zhangsan]
c. Zhangsan zhidao Lisi yiding ban-de-dong ziji de gangqin. [ziji = Lisi]
d. Zhangsan zhidao Lisi yiding ban-de-dong ziji de gangqin. [ziji = Zhangsan]

(9)	a.	[+number phrase]	[+local binding]	0/4 yes
	b.	[+number phrase]	[-local binding]	4/4 yes
	c.	[-number phrase]	[+local binding]	4/4 yes
	d.	[-number phrase]	[-local binding]	3/4 yes

Using the functions in the appendix, these (imaginary) results show that the interaction between the factors [number phrase] and [local binding] is statistically significant ($p < .05$), providing (imaginary) support for one of the claims in Li (1998). This is the smallest experiment of this type capable of showing an interaction where the [+F+G] sentence is expected to be bad but all of the other sentences should be good. If the predicted interaction were instead that both [+F+G] and [-F-G] interactions should be bad, a significant effect is possible with as few as three speakers per sentence (for a total of twelve). The smallest experiment capable of showing two significant effects (e.g. both factors, or one factor and the interaction) requires six speakers per sentence (24 total), and to test if all three factors are significant, we require a minimum of nine speakers per sentence (36 total) (see Myers et al. 2007).

A simple extension of the exact McNemar test to two-factor within-groups experiments is also possible if we recognize that a speaker giving binary judgments to four sentences can produce any one of sixteen (= 2^4) possible judgment patterns. For factor F (and similarly for G and FG), only ten of the sixteen possible outcomes are discordant, with five showing a greater number of yes judgments for [+F] than [-F] and five showing the reverse. Since by chance there should be equal numbers of the two types of discordant responses in a sample of judgment quadruples grouped by speakers, we can model the chance situation by flipping a coin, just as with paired judgments. Moreover, the design structure means that results for F, G, and FG are independent of each other, so we can test each using the exact McNemar test. For example, if we ask six speakers (the minimum number capable of showing a significant effect) to judge all four sentences in (8), and all accept (8ab) and all reject (8cd), this would indicate a significant main effect for [number phrase] (nondirectional $p < .05$). Unlike the other techniques described so far, however, this simple extension of the exact McNemar test isn't the most sensitive possible, since it neglects the fact that judgment patterns in a two-factor experiment can be discordant to different degrees (e.g. a speaker may accept (8a), (8b) and

(8c) but reject (8d), which represents a weaker [number phrase] effect); a more complex method would be needed to take the different effect strengths into account.

The first lesson of this section is that a two-factor experiment can be thought of as representing a kind of regression problem, with three orthogonal predictors. This way of looking at the situation is crucial to understanding the more powerful technique described in the next section.

The second, more important, lesson is that once we recognize that binary judgments are worthy of respect, as is the desire for judgment experiments that are as small as possible, we may be led to adopt previously unfamiliar statistical techniques better suited to our goals. It doesn't make sense to force all judgment experiments into the mold defined by statistical tools like analysis of variance (ANOVA), simply because they are commonly used in research domains where, unlike theoretical syntax, large-scale experiments with continuous-valued measurements are the norm.

3.3 A generalized method

Though it is not unreasonable to run experiments testing one set of sentences on multiple speakers, as described in the previous sections, linguists recognize that judgments can vary in theoretically uninteresting ways across both sentences and speakers. Hence the family of small-scale experiments should include those where we test some number of sentences on some number of speakers, so that both sources of noise can be extracted from the statistical analysis. Moreover, given that judgments may shift over the course of the experiment, we should consider factoring out the effects of presentation order as well. This is especially important if the experiment doesn't use counterbalancing or fillers, techniques for dealing with cross-sentence interference that complicate the experimental methodology and thus begin to move out of the realm of the small-scale experiment.

The first challenge, variation across both speakers and sentences, has traditionally been handled in psycholinguistics by running separate analyses for each. When this procedure was first proposed by Clark (1973), the point was to bring the two analyses back together again in a single analysis at the end. Today most psycholinguistics papers describe only the separate by-participants and by-items analyses, though Raaijmakers, Schrijnemakers, and Gremmen (1999) have sparked a renewal of interest in Clark's original method.

In the meantime, however, statisticians developed a much more powerful method for handling the problem of participants and items, the so-called mixed-effects regression model (Pinheiro and Bates 2000). Though not yet dominant in psycholinguistics, it is likely to take over the field eventually given the advocacy of textbooks like Baayen (2008). Mixed-effects models are so called because they mix the "fixed" factors that define the experimental design (e.g. the presence/absence of *you*) with one or more "random" factors (e.g. speakers and/or sentences) in a single equation. This not only increases statistical power, but also makes it possible to compare by-speakers and by-speakers-and-sentences models to determine whether there is any advantage to including the second random factor. Thus it sometimes may turn out that a simple by-speakers analysis is sufficient, despite current psycholinguistic practice (Raaijmakers et al. 1999 draw the same conclusion from within the earlier Clark framework).

When the observations are binary, as in small-scale judgment experiments, the relevant variety is called mixed-effects logistic regression (Agresti 2002, Baayen 2008, Jaeger 2008, Moreton forthcoming). Logistic regression is the statistical workhorse of variationist sociolinguistics (see e.g. Mendoza-Denton, Hay, and Jannedy 2003), and by putting it into a mixed-effects context, we can analyze binary judgments across both speakers and sentences in a single statistical model. By contrast, the Clark (1973) method requires prior averaging, which isn't appropriate for binary data (see Baayen 2008, Jaeger 2008).

Mixed-effects logistic regression really has only one limitation in relation to small-scale experimentation, and that is that unlike the statistical tests described earlier, it is more accurate with larger samples. The previous tests are called exact tests, since the *p* values directly reflect the probabilities associated with the space of possible events; that is, we count the number of "winning" cases and divide by the number of logically possible cases. Exact tests can be computationally intensive, which is why procedures like exact logistic regression are restricted to special-purpose software. Hence it is convenient to approximate the exact *p* values with formulas that become more reliable the larger the sample size, and this is what is typically done with mixed-effects models. On top of this, logistic regression involves a type of estimation algorithm that crashes, ironically enough, when correlations are too "perfect" (Albert and Anderson 1984), and "perfect" correlations are more likely to happen with very small samples.

Fortunately, because mixed-effects logistic regression deals with the raw data, not averages across speakers or sentences, it defines sample size in terms of the total number of observations, which can be sufficiently large even for very small experiments. As a demonstration of this, I simulated 10,000 judgment experiments with one to five "speakers" and one to five quadruples of "sentences", and tested the effect two binary factors, their interaction, and the order of sentence presentation on randomly generated "judgments" using by-speaker mixed-effects logistic regression. Since the data are random, we expect the probability of getting a "significant" result ($p < .05$) for any given factor or interaction to be .05 (this follows from the definition of *p* values). The simulations showed that when the number of observations was below 32 (e.g. four speakers judging two quadruples, or two speakers judging four quadruples), the proportion of "significant" results was too small (.03 or below). This is related to the minimal sample size limitations we saw with the exact tests, but also to the algorithm crashing with "perfect" correlations. Above this minimum, however, "significant" results are found at a rate between .05 and .06, only moderately above the ideal.

Of course larger experiments give even more reliable *p* values, but the improvement is not dramatic. For example, for a two-factor experiment with ten speakers and ten sets (400 data points), running 10,000 simulated experiments yield a "significance" rate of .051. Hence it seems that small-scale judgment experiments need no more than a total of 32 judgments to provide reasonably reliable *p* values.

To run a mixed-effects logistic regression on a small-scale judgment experiment, the one or two fixed factors are first coded as described in section 3.2, so that [+F] = 1 and [-F] = -1, and interactions are the products between factors. Judgments are coded as yes = 1 and no = 0. Speakers and sentences are given identification labels so that observations can be grouped accordingly (i.e. this observation from this speaker judging this sentence). The result of a mixed-effects logistic regression analysis is a table showing the (nondirectional) *p* values associated with each factor and interaction, along with a coefficient indicating the size and direction of each effect.

The fact that mixed-effects logistic regression analyzes raw observations, rather than averages as in traditional by-participant and by-item ANOVAs, means that it is well suited for dealing with the challenge of order effects. Randomization of presentation order is one technical feature of full-fledged psycholinguistic experimentation that is worth maintaining in even in small-scale experimentation, since it automatically factors out nuisance effects like fatigue, overpractice, and priming of later sentences by earlier ones. It is also very simple to do, as explained in section 3.5. Since presentation order must be generated for each speaker anyway, it is no trouble to keep a record of it after the experiment is done, coded as an ordinal number (1 for the first sentence presented, 2 for the second, and so on). Presentation order can thus be treated as an ordinal fixed factor along with the fixed binary factor(s).

But we can go further than this, by looking at the interactions of the binary factors (and their interaction) with order. An interaction with order implies a change in the strength of the factor, that is, the difference between the two values of the factor, over the course of the experiment. If this change involves the judgment contrast getting weaker, this is called syntactic satiation (Snyder 2000). The reverse is also possible, where a judgment contrast gets stronger; Ko (2007) reports several examples of such "anti-satiation" cases, which apparently arise because speakers learn to overcome their initial parsing difficulties. Interactions with order are easy to include within a mixed-effects logistic regression model, and even if the researcher is not interested in satiation per se, factoring out order interactions can help make effects of the main factors stand out more clearly (see Myers 2007ab for examples).

Estimating *p* values for mixed-effects logistic regression is much more computationally intensive than the techniques described earlier, and Baayen, Davidson, and Bates (2008) claim that the lme4 package (Bates and Sarkar 2007) in the free statistics program R (R Development Core Team 2008) is currently the only software tool capable of handling multiple random variables in mixed effect models. Fortunately, its implementation in R means that it is readily available to any researcher willing to take the bit of extra effort needed to learn how to use it (see Appendix A6 gives more information).

3.4 Automation

Though the procedures for proper experimental design and statistical analysis are eminently learnable, they are even easier to use if the most difficult steps are automated. In this section I describe MiniJudge, a free software tool for doing just that (Myers 2008a, Chen, Yang, and Myers 2007; for applications see Myers 2007ab, Myers 2008b, Ko 2007, Lawrence 2007). MiniJudge is designed not only to simplify the statistical analysis (it outputs a simplified results summary and a graph), but also the selection of experimental materials and the creation and distribution of judgment surveys. MiniJudge currently exists in two implementations: MiniJudgeJS, written in HTML and JavaScript, and MiniJudgeJava, written in Java, and R is used for statistical analysis.

As we have seen, theoretical syntacticians are accustomed to generating well-matched sentence pairs (and sometimes quadruples). As we have also seen, however, testing only one set makes it impossible to generalize to other sentences. Generating variants on a set of contrasting sentences is not always easy, since it requires changing lexical content and other irrelevant aspects while maintaining core sentence structure. Cowart (1997) describes a semi-automated procedure for doing this in a spreadsheet program, by replacing strings of words in one sentence for another, and MiniJudge builds on this procedure.

After the researcher has chosen the one or two binary factors defining the experimental design, a so-called prototype set of two (or four) sentences is entered into MiniJudge, representing the sort of core examples typically cited in a syntax paper (like those in Li 1998). To help generate variants, MiniJudge then divides the prototype sentences into the largest possible substrings of words. For example, suppose the examples in (10) were used as the prototype set in a one-factor experiment on complex NP islands in English. MiniJudge extracts the following segments: *who does John believe*, *that Bill saw*, *the claim*. The string *who does John believe* is segmented out because its left context includes a sentence boundary and its right context can vary (sometimes *that* and sometimes *the*). The other segments are extracted for similar reasons. Note that segmentation is done solely on the basis of the prototype sentences as word strings; MiniJudge has no linguistic knowledge. This process works just as well for orthographies, like that of Chinese, that do not mark word boundaries.

(10) a. Who does John believe that Bill saw?
b. Who does John believe the claim that Bill saw?

Now, rather than having to invent new sentence sets word by word, risking mismatches, the researcher only needs to choose syntactically equivalent substitutes for the three prototype segments. For example, *who does John believe* may be replaced with *what did Mary hear*, *that Bill saw* with *that Jane ate*, and *the claim* with *a rumor*. MiniJudge then fits these substitutes into the same positions occupied by the originals, generating the new sentence set in (11). MiniJudge's lack of linguistic knowledge means that the result may not always be quite what was desired, in which case the researcher has to tweak the new sentence sets a bit by hand, but this is still simpler than creating them from scratch.

(11) a. What did Mary hear that Jane ate?
b. What did Mary hear a rumor that Jane ate?

MiniJudge, in its current implementations, is a fully functional tool that has been used successfully to test a variety of linguistic hypotheses. The most important purpose of MiniJudge, however, is that it represents proof of concept: software designed for small-scale judgment experimentation is possible and desirable. All of the computer code in MiniJudge is open-source, and programmers are welcome to borrow whatever they like, or to reject all of it and start over again. The ultimate goal is to provide as practical and non-intimidating a tool for theoretical syntacticians as possible, and competing projects along the same lines can only help speed up achievement of this goal.

3.5 Application to Li (1998)

To round out this tutorial, recall that Li (1998) contains several properly designed informal one- or two-factor experiments. The most important missing element was explicit quantification, or more precisely, statistical testing of the claim that Li's judgments are representative of all Chinese speakers. Testing whether the sentences are also representative of the Chinese sentence inventory would require creating new materials, and although this task would be greatly simplified with MiniJudge, the goal here wasn't to retest Li's claims with new experiments so much as to add the element of quantification to her already well-designed study.

We thus extracted all of the Chinese examples cited in Li (1998), retyped them in Chinese characters, along with any necessary context, such as the intended interpretation of reflexives, as illustrated in (8) above. The four claims associated with Li's one-factor designs and the two claims associated with her two-factor designs are listed in the left side of Table 1, along with the 24 sentences (including contextual variants) she used to test them.

[INSERT TABLE 1 HERE]

The remaining Chinese sentences in Li's paper served as fillers, giving a total of 38 sentences per survey. Each survey presented the sentences in a different random order, which was accomplished by entering the sentences into a spread-sheet program, adding a column of random numbers on the left (using the program's built-in random number generating function), and then ordering both columns by the random numbers. The surveys were printed separately, along with instructions to judge each sentence as good (coded as 1) or bad (coded as 0), depending on whether they seemed Chinese-like or likely to be used by a Chinese-speaking person. Each sentence had to be judged one at a time, without going back or skipping. The surveys were then completed by twenty-four native speakers (college students) of the same variety of Mandarin Chinese spoken by Li (Taiwan Mandarin). An additional six speakers were tested, but five had had previous experience with linguistic analysis, and the sixth (the last linguistically naive speaker to be run) was dropped to make the number of speakers divisible by four (for reasons explained shortly).

Raw judgment patterns for most of the six claims trended in the direction claimed by Li (1998), with a greater number of yes judgments for sentences claimed to be grammatical in comparison to the appropriate controls. This can be seen from the signs of the bolded mixed-effects coefficients shown in the right side of Table 1, which for the most part are positive when the construction defined by the positive values of the factors should be acceptable and negative when the construction should be unacceptable (the one exception is discussed below).

In order to compare the relative strengths and weaknesses of the statistical tests discussed in this paper, the results were analyzed with all of them. The mixed-effects analyses were only run by-speaker and ignored order, since there were never enough items per factor to make it possible to include order as a fixed factor or sentences as a random factor. The exact tests are not designed to handle multiple speakers and multiple sentences at the same time, so redundant sentence sets were dropped; in practice this meant that only the first pair of sentences in one-factor designs were included in the analysis. The exact between-groups tests involved first dividing the speakers into two groups (for one-factor designs) or four groups (for two-factor designs), so that judgments for each sentence type came from a different group, as required by this type of analysis. See the appendix for further information on the data and the analyses.

The resulting *p* values (and regression coefficients for mixed-effects analyses) are shown in the right side of Table 1. Overall, Li's claims did quite well, with four out of six claims statistically significant in the most powerful of the tests (mixed-effects logistic regression). A binomial test (with a "coin" that comes up heads only 5% of the time) shows that the chance probability of an outcome at least as good as this is less than .0002. The significant effects include both of the predicted interactions, despite the fact that these involved subtle semantic judgments (binding of reflexives and scope). Aside from their pedagogical value, then, these results confirm once again that linguistically naive speakers are capable of providing primary intuitions relevant to the testing of theoretically interesting syntactic hypotheses.

The two claims that failed to reach statistical significance were, first, the claim that number phrases in subject position improve in answers to *how many* questions (cf. (3) above), and second, the claim that existential *you* permits number phrases to bind pronouns. The latter pattern even trended in the opposite direction from Li's claim (i.e. adding *you* slightly decreased yes responses instead of increasing them as predicted). As with any null result, from these data alone it is impossible to know what went wrong here, though we may speculate. Since the *how many* effect trended in the right direction, it may simply be that the effect is too subtle to detect with so few observations, and more sentence sets should be added. The failure to detect the pronoun binding effect might possibly relate to the complexity of the sentences that Li used to test it, where the pronoun was located inside a center-embedded relative clause. This time adding more sentence sets may help both by giving linguistically naive speakers sufficient parsing practice, and by providing enough experimental trials to make it possible to factor out any interaction with presentation order that may be obscuring the main effect (see discussion of satiation above in section 3.3).

Methodologically, Table 1 shows that mixed-effects logistic regression was the most sensitive, with *p* values consistently lower than those of the within-groups exact tests, which generally had lower *p* values than the between-groups exact tests. This difference relates partially to differences in the number of observations available for the three types of tests in this omnibus experiment. The number of analyzable observations was greatest for mixed-effects modeling (96, all of the data), less for within-groups exact tests (half of the data for one-factor designs), and least of all for between-groups exact tests (only 24, the number of speakers). Combined with the demonstration in section 3.3 that mixed-effects logistic regression is reliable even for samples with as few as 32 observations, this test earns its status as the default. Nevertheless, as was also demonstrated earlier, when the number of observations go below this, or the logistic regression algorithm crashes, exact tests become the only option.

Although twenty-four speakers were tested, the significant patterns in Table 1 could have been detected with fewer. For example, consider the first claim noted in Li (1998), that number phrases in subject position require existential *you* (aside from the other factors addressed later in Li's paper). When the extended exact McNemar test ($p < .05$) is rerun on randomly chosen subsets of the speakers (throwing out data from all other speakers), we find that the effect probably would have been detected with as few as eight speakers (more precisely, it was detected in more than half of the 10,000 resampled data sets with this many speakers). However, this particular syntactic effect is apparently very robust; not only was it well known prior to Li (1998), but in this experiment it was detected by all three of the statistical tests. By contrast, the weakest of the significant effects in this experiment is the interaction associated with reflexive binding, significant only by mixed-effects logistic regression. Nevertheless, even for this effect, resampling shows that eighteen speakers would have been sufficient to detect it more likely than not (a "savings" of six speakers).

Despite its success in providing objective empirical evidence for many of the claims in Li (1998), using relatively few speakers and relatively simple methods, this small-scale experiment was admittedly more difficult to run than Li's original informal judgment experiments. Even with automation, it takes a few hours to create and analyze a set of surveys, and researchers need to have the means (and, usually, permission from an ethics board) to recruit native speakers of the language of interest. Nevertheless, given careful design and proper statistics, very few speakers and sentences need to be tested, and with a bit of practice the process can become second nature (a point also emphasized by Cowart 1997 for larger-scaled experiments).

4. Conclusions

The informal collection of judgments involves a genuinely experimental methodology. Theoretical syntacticians understand the importance of factorial designs, well matched materials, and controls, and they do attempt to test multiple sentences, and, if necessary, multiple speakers, in order to confirm that claimed generalizations actually do generalize. This is why the judgment data cited in the traditional syntax literature provide a reasonably solid empirical basis for theoretical analysis.

Yet this empirical basis could easily be strengthened by applying traditional methodologies with a bit more rigor. In particular, factorial designs are used, but they are not applied consistently, particularly when more than one factor is involved. Applying this extra rigor is easy because it builds on the traditional methods themselves. Small-scale experiments, with results tested by simple but surprisingly powerful statistical techniques, can be completed quickly, since very few speakers and sentences need be tested, and judgments can be of the familiar binary type. Syntacticians concerned about the reliability of their judgments shouldn't think that their only alternative is to run a full-fledged formal judgment experiment of the sort described in most of the experimental syntax literature. While such experiments may define the gold standard, in most cases a quick-and-dirty small-scale experiment, if designed and analyzed properly, may be sufficient to satisfy most skeptics.

Acknowledgments

My thanks to Yu-guang Ko for helping with bibliometric counts and running the small-scale experiment. MiniJudgeJS, MiniJudgeJava, and the R code for the generalization of Fisher's test are co-copyrighted by National Chung Cheng University. MiniJudgeJava was written by Tsung-ying Chen and Chen-Tsung Yang. Research for this paper was supported by National Science Council (Taiwan) grants NSC 94-2411-H-194-018, NSC 95-2411-H-194-005, and NSC 96-2411-H-194-002. All errors are solely my responsibility.

Table 1. Results from by-speaker analyses of the judgment contrasts claimed in Li (1998)

			Mixed-effects logistic regression			Exact within-groups tests			Exact between-groups tests		
Factors	Judgment claim	Examples in Li (1998)	F	G	FG	F	G	FG	F	G	FG
F = [*you*]	The absence of *you* ("have") is disallowed with number phrases	(1) vs. (3) (2) vs. (4)	**.000** **2.2**			**.000**			**.009**		
F = [how many]	*You*-less number phrases improve in a *how many* context	(1) vs. fn. 3 (2) vs. fn. 3	**.390** **0.18**			**1**			**1**		
F = [*you*]	*You* is disallowed with quantity expressions like *gou* ("enough")	(6) vs. (16) (8) vs. (17a)	**.001** **-0.9**			**.021**			**.069**		
F = [*you*]	*You* permits number phrases to bind pronouns	(20a) vs. (20b) (21a) vs. (21b)	**.238** **-0.3**			**1**			**.317**		
F = [number] G = [binding]	Number phrases can't bind reflexives	(22a): two meanings (22b): two meanings	.543 -0.2	.000 -1.5	**.014** **-0.7**	1	.000	**1**	.425	.425	**.103**
F = [*you*] G = [scope]	*You* allows number phrases to scope over lower ones	(9): two meanings (23a): two meanings	.010 -0.8	.001 -1.0	**.000** **1.2**	1	.000	**1**	.112	1	**1**

NOTE. Example numbers are those used in Li (1998). In the cells in the nine columns on the right, the first values are *p* values (*p* < .05 indicates statistical significance). The second values in the mixed-effects results are regression coefficients. Bolding indicates the values relevant to testing the hypotheses in Li (1998).

References

Adli, A., 2005. Gradedness and consistency in grammaticality judgments. In: Kepser, S., Reis, M. (Eds.), Linguistic Evidence: Empirical, Theoretical and Computational Perspectives. Mouton de Gruyter, The Hague, pp. 7-25.

Agresti, A. 2002. Categorical Data Analysis (second edition). Wiley-Interscience, Hoboken, NJ.

Albert, A., Anderson, J., 1984. On the existence of maximum likelihood estimates in logistic\ regression models. Biometrika 71(1), 1-10.

Aoun, J., Li, Y.-H. A., 2003. Essays on the Representational and Derivational Nature of Grammar: The Diversity of Wh-constructions. MIT Press, Cambridge, MA.

Baayen, R. H. 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics. Cambridge University Press, Cambridge, UK.

Baayen, R. H., Davidson, D. J., Bates, D. M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language.

Bard, E. G., Robertson, D., Sorace, A., 1996. Magnitude estimation of linguistic acceptability. Language 72 (1), 32-68.

Bates, D., Sarkar, D., 2007. lme4: Linear mixed-effects models using S4 classes. R package.

Bernstein, J. B., Cowart, W., McDaniel, D., 1999. Bare singular effects in genitive constructions. Linguistic Inquiry 30 (3), 493-502.

Chen, L., & Pan, N. 2003. The categorical status of finite complements of xiangxin 'believe' and renwei 'think' in Chinese. In: Lin, Y.-H. (Ed.), Proceedings of the Fifteenth North American Conference on Chinese Linguistics, pp. 45-53.

Chen, T.-Y., Yang, C.-T., Myers, J., 2007. MiniJudgeJava (Version 0.9.9) [Computer software]. Accessible at http://www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm.

Chomsky, N., 1957. Syntactic Structures. Mouton, The Hague.

Chomsky, N., 1965. Aspects of the Theory of Syntax. MIT Press, Cambridge, MA.

Chomsky, N., Lasnik, H., 1977. Filters and control. Linguistic Inquiry 8, 425–508.

Clark, H., 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. Journal of Verbal Learning and Verbal Behavior 12, 335-359.

Clifton, Jr., C., Fanselow, G., Frazier, L., 2006. Amnestying superiority violations: Processing multiple questions. Linguistic Inquiry 37 (1), 51-68.

Cowart, W., 1997. Experimental Syntax: Applying Objective Methods to Sentence Judgments. Sage Publications, London.

Crain, S., Thornton, R., 1998. Investigations in Universal Grammar: A Guide to Experiments in the Acquisition of Syntax and Semantics. MIT Press, Cambridge, MA.

Derr, R. E., 2000. Performing exact logistic regression with the SAS® System. SUGI 25 Proceedings (Paper P254-25). Available at http://www2.sas.com/proceedings/sugi25/25/st/25p254.pdf and mirrors.

Featherston, S., 2005a. That-trace in German. Lingua 115 (9), 1277-1302.

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

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

Fisher, R. A., 1926. The arrangement of field experiments. Journal of the Ministry of Agriculture 33, 503-513.

Fisher, R. A., 1935. The mathematics of a lady tasting tea (originally a passage in Design of experiments). Reprinted and retitled 1956 in: Newman, J. R. (Ed.), The World of Mathematics. Simon and Shuster, New York, pp. 1512-1521.

Hill, A. A., 1961. Grammaticality. Word 17 (1), 1-10.

Huang, J., 1982. Logical Relations in Chinese and the Theory of Grammar. Doctoral dissertation, MIT, Cambridge, MA.

Ishii, T., 2006. A nonuniform analysis of overt wh-movement. Linguistic Inquiry 37 (1), 155-167.

Jaeger, T. F., 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. Journal of Memory and Language.

Johnson, K., 2008. Quantitative methods in linguistics. Blackwell, Oxford, UK.

Kirk, R. E., 1995. Experimental Design: Procedures for the Behavioral Sciences. Brooks/Cole, New York.

Ko, Y.-G., 2007. Grammaticality and Parsability in Mandarin Syntactic Judgment Experiments. Master's thesis, National Chung Cheng University, Chiayi, Taiwan.

Labov, W., 1975. Empirical foundations of linguistic theory. In: Austerlitz, R. (Ed.), The Scope of American Linguistics. Peter de Ridder, Lisse, pp. 77-133.

Lawrence, D., 2007. Using MiniJudge to test sentence acceptability. Talk presented at Thinking Matters Conference, University of Southern Maine.

Lee, H. T., 1986. Studies on quantification in Chinese. Doctoral dissertation, UCLA.

Li, Y.-H. A., 1998. Argument determiner phrases and number phrases. Linguistic Inquiry 29 (4), 693-702.

Li, Y.-H. A., 2004. Linguistics as an empirical science: Movement structures as a case study. Talk presented at National Chung Cheng University, Taiwan.

Manning, C. D., Schütze, H., 1999. Foundations of statistical natural language processing. MIT Press, Cambridge, MA.

McNemar, Q., 1947. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika 12 (2), 153-157.

Mendoza-Denton, N., Hay, J., Jannedy, S., 2003. Probabilistic sociolinguistics: Beyond variable rules. In: Bod, R., Hay, J., Jannedy, S. (Eds.), Probabilistic Linguistics. MIT

Press, Cambridge, MA, pp. 97-138.

Moreton, E., Forthcoming. Analytic bias and phonological typology. Phonology.

Myers, J., 2007a. MiniJudge: Software for small-scale experimental syntax. International Journal of Computational Linguistics and Chinese Language Processing 12 (2), 175-194.

Myers, J., 2007b. Generative morphology as psycholinguistics. In G. Jarema & G. Libben (Eds.), The mental lexicon: Core perspectives (pp. 105-128). Amsterdam: Elsevier.

Myers, J., 2008a. MiniJudgeJS (Version 1.1) [Computer software]. Accessible via http://www.ccunix.ccu.edu.tw/~lngproc/MiniJudge.htm

Myers, J., 2008b. Automated collection and analysis of phonological data. Talk presented at the International Conference on Linguistic Evidence 2008, Tübingen, Germany.

Myers, J., Huang, S.-F., Tsay, J. 2007. Exact conditional inference for two-way randomized Bernoulli experiments. Journal of Statistical Software 21, Code Snippet 1, 2007-09-02.

Nickerson, R. S., 1998. Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology 2 (2), 175-220.

Nisbett, R. E., Wilson, T. D., 1977. Telling more than we can know: verbal reports on mental processes. Psychological Review 84 (3), 231-259.

Ou, T., 2006. Suo relative clauses in Mandarin Chinese. Master's thesis, National Chung Cheng University, Chiayi, Taiwan.

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

Pinheiro, J. C., Bates, D. M., 2000. Mixed-Effects Models in S and S-Plus. Springer, Berlin.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., Gremmen, F., 1999. How to deal with "the language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. Journal of Memory and Language 41, 416-426.

Schmitz, H.-C., Schröder, B., 2002. On focus and VP-deletion. Snippets 5, 16-17.

Schütze, C. T., 1996. The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology. University of Chicago Press, Chicago.

Schütze, C. T., 2001. On the nature of default case. Syntax 4(3), 205-238.

Schütze, C. T., 2005. Thinking about what we are asking speakers to do. In: Kepser, S., Reis, M. (Eds.), Linguistic Evidence: Empirical, Theoretical and Computational Perspectives. Mouton de Gruyter, The Hague, pp. 457-484.

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

Snyder, W., 2000. An experimental investigation of syntactic satiation effects. Linguistic Inquiry 31, 575-582.

Soh, H. L., 2005. Wh-in-situ in Mandarin Chinese. Linguistic Inquiry 36 (1), 143-155.

Sorace, A., Keller, F., 2005. Gradience in linguistic data. Lingua 115, 1497-1524.

Sprouse, J., 2007. Continuous acceptability, categorical grammaticality, and experimental syntax. Biolinguistics 1, 123-134.

Tang, T. C., 1984. Hanyu Cifa Jufa Lunji [Essays on Chinese Morphology and Syntax]. Student Books Co., Taipei, Taiwan.

Wasow, T., Arnold, J., 2005. Intuitions in linguistic argumentation. Lingua 115, 1481–1496.

Weskott, T., Fanselow, G., 2008. Different measures of linguistic acceptability: Not so different after all? Presented at the International Conference on Linguistic Evidence 2008, Tübingen, Germany, January 31 - February 2.

Whitman, P. N., 2002. Category Neutrality: A Type-Logical Investigation. Ohio State University PhD thesis.

Woods, A., Fletcher, P., Hughes, A., 1986. Statistics in language studies. Cambridge University Press, Cambridge, UK.

Xu, L., 1990. Remarks on LF Movement in Chinese questions. Linguistics 28, 355-382.

Xu, L., 1996. Construction and destruction of theories by data: A case study. Chicago Linguistics Society 32, 107-118.

Appendix: Analyzing small-scale judgment experiments with R

A1. R basics.

The latest version of R for Windows, Macintosh, or Linux may be downloaded for free from http://cran.r-project.org/mirrors.html. Follow the online instructions to install it. The R scripts for the special functions below, as well as the data files for the experiment and the simulations mentioned in sections 3.3 and 3.5, are available online at http://www.ccunix.ccu.edu.tw/~lngproc/SmallScale.htm. In the sample code below, underlining indicates the portions that are not part of the R language itself and which should be changed according to the data being analyzed.

A2. The exact McNemar test.

Let Pattern and Exception represent the numbers of the two types of outcomes (discordant judgment pairs). Then the nondirectional *p* value is computed with following R command:

(A1) min(1,2*pbinom(min(Pattern,Exception),sum(Pattern,Exception),0.5))

Note that "e-" in a *p* value usually means that it is very small, since the value following it relates to the number of zeroes after the decimal point. For example, 2.6e-4 = 2.6 × $(1/10)^4$ = 0.00026.

A3. Fisher's exact test.

Let a, b, c, d represent the values of the cells arranged as in the following table, where a vs. b is the yes vs. no contrast for factor value [+F] and *c* vs. *d* is the yes vs. no contrast for factor value [-F].

(A2) Setting up Fisher's exact test

	[+F]	[-F]
yes	a	c
no	b	d

Then the nondirectional *p* value for F is computed with the following R command:

(A3) fisher.test(cbind(c(a,b),c(c,d)))$p.value

A4. An extension of Fisher's exact test.

Suppose N is the number of each speaker group, a, b, c, d are the "yes" counts of the cells arranged as in the following table, and "F" and "G" are the names of the factors.

(A4) Setting up the extension of Fisher's exact test

	[+F]	[-F]
[+G]	a	c
[-G]	b	d

An R script for creating the function ext.fisher is available at http://www.ccunix.ccu.edu.tw/~lngproc/SmallScale.htm. After this script has been pasted into R, the following command will compute the nondirectional *p* values for the two factors and their interaction (note that the quotation marks are required). Note that the current version of this algorithm requires that the maximum number of yes judgments in each cell be identical.

(A5) ext.fisher(cbind(c(a,b),c(c,d)),N,"F","G")[,3]

If no *p* value exists (since there is no alternative possible arrangement of data to compare with the observed data), the function gives "NA" (not available).

A5. A function for running exact tasks

To make it easier to run the above exact tests, the webpage http://www.ccunix.ccu.edu.tw/~lngproc/SmallScale.htm also has an R script creating the omnibus function small.exp, which can automatically choose among the above tests based on the number of factors (one vs. two) and grouping (between-groups vs. within-groups); it then outputs nondirectional *p* values for the factor(s) and any interaction. Schematic examples for the four logically possible applications are follows.

(A6) One-factor, between-groups: small.exp(Judgment, Factor1)
One-factor, within-groups: small.exp(Judgment, Factor1, group=Speaker)
Two-factor, between-groups: small.exp(Judgment, Factor1, Factor2)
Two-factor, within-groups: small.exp(Judgment, Factor1, Factor2, group=Speaker)

This function assumes that the data are organized in a file as described in section A6 below, but remember that the above exact tests cannot handle sentences as a random variable, and each one makes different assumptions about grouping. For example, for a one-factor, between-groups test, the test assumes that half of the speakers provide judgments for [+F] sentences and the other half judgments for [-F] sentences. Thus if you give it data where each speaker judges both sentence types, the function will crash unless you specify that grouping is by speaker.

A6. Mixed-effects logistic regression.

To analyze judgment data with mixed-effects logistic regression, you must first create a data file (examples are available at http://www.ccunix.ccu.edu.tw/~lngproc/SmallScale.htm). Your data file, which can be created in a spreadsheet program, should have columns labeled Speaker, Item, F (and G), Order, and Judgment, where F and G represent the actual names of the factors (factor names in R cannot contain spaces or punctuation marks other than "." or "_").

Below the header line, each row in the data file represents a unique observation (data point). The Speaker and Sentence columns show unique identification numbers for each speaker and each sentence, respectively. Each sentence is coded for the factor(s) it represents in the F and G columns, where 1 = [+F] (or [+G]) and -1 = [-F] (or [-G]). The Order column shows the position in the list of sentences for the given speaker and sentence (this column isn't necessary if you have no interest in order effects). The Judgment column shows judgments, where 1 = acceptable and 0 = unacceptable.

Name and save the data file as a tab-delimited text file (e.g. "expdata.txt"). This can be done by copying and pasting from the spreadsheet into a simple text editor (e.g. Microsoft Notepad). Then load the data into R using the following command (the quotation marks are necessary); the *attach* command allows the column names to be treated as variable names:

```
(A7) dataset = read.table("expdata.txt",T)
     attach(dataset)
```

To install lme4, the special package needed to run mixed-effects analyses in R, either choose *Packages... > Install package(s)...* in the R menu, or type the following command into R's main window (your computer must be connected to the internet):

```
(A8) install.packages("lme4", dependencies = TRUE)
```

After lme4 is installed on your hard drive, you will still need to load it into memory each time you run mixed-effects analyses in R. You can do this either with *Packages... > Load package...* in the R menu, or with the following command:

```
(A9) library(lme4)
```

The general structure of a command creating a mixed-effects logistic regression is as follows. Give each analysis you run a different name so that you can compare them.

```
(A10)    mixlog = lmer(Judgment ~ FixedFactors + RandomFactors, family = "binomial",
         data = dataset)
```

The elements labeled "FixedFactors" and "RandomFactors" should be filled in differently depending on what you want to test, as described in tables (A10) and (A11).

(A11) Choosing fixed factors

Analysis	FixedFactors
One factor, ignore order	F
One factor, test simple order effects	F + Order
One factor, test for (anti)satiation	F * Order
Two factors, ignore order	F * G
Two factors, test simple order effects	F * G + Order
Two factors, test for (anti)satiation	F * G * Order

(A12) Choosing random factors

Analysis	RandomFactors
By speakers only	(1|Speaker)
By speakers and sentences	(1|Speaker) + (1|Item)

For example, if you want to run a by-speakers-only analysis for an experiment with two factors where you don't care about order effects, you would use the following command.

```
(A13) mixlog = lmer(Judgment ~ F*G + (1|Speaker), family = "binomial", data = dataset)
```

To determine whether a by-speaker-and-sentence analysis does a better job than a by-speaker-only analysis (without changing the FixedFactors portion of the model formula), use the following command (where mixlog1 and mixlog2 represent the simpler and more complex analyses, respectively). If $p < .05$, the sentences significantly affect how the factors influence judgments, so the more complex analysis is preferred.

```
(A14) anova(mixlog1, mixlog2)[2,7]
```

To see the *p* values associated with the factors in a mixed-effects logistic regression analysis, simply type the name of the analysis, or use the following command for an abbreviated report.

(A15) summary(mixlog)@coefs

For further information on how to run and interpret mixed-effects logistic regression analyses (run using a method called generalized linear mixed-effects modeling, or GLMM) for small-scale judgment experiments, visit the MiniJudge help page at http://www.ccunix.ccu.edu.tw/~lngproc/MJInfo.htm.