REACTION TIME AS A SUPPLEMENT TO GRAMMATICALITY JUDGEMENTS IN THE INVESTIGATION OF SECOND LANGUAGE LEARNERS' COMPETENCE

ROBERT BLEY-VROMAN DEBORAH MASTERSON
University of Hawai'i

Introduction

It is frequently important for the language acquisition researcher to determine whether a particular sentence is licensed by the developing internal grammar of the learner—that is,whether a sentence is grammatical for the learner.

Many kinds of SLA research have been interested in grammaticality. For example, researchers studying how learners master the article system of English, and what uses of the articles are more difficult than others, will investigate grammaticality. Researchers interested in global proficiency may want to determine the degree to which the learner's grammar corresponds to the target language grammar—perhaps correlating a general correctness score with some other variable of interest.

However, the research program with which this paper is concerned is not really interested in the grammaticality of particular examples per se, and certainly not in obtaining correctness scores or developing reliable proficiency measures; rather, the research is directed toward ascertaining something quite specific about the hypothesized internal grammatical system itself. It is centrally concerned with the structure of that system, not with the learner's language output.

In this research program, some general conception of the nature of language acquisition and of the human language faculty leads the researcher to a particular hypothesis about the character of the internal grammar, and this in turn leads to a prediction that a certain string of words will or will not be a possible sentence of the language. To take a concrete example, Bley-Vroman, Felix, & Ioup (1988) ask a general question about the character of adult second

University of Hawai'i Working Papers in ESL, Vol. 8, No. 2, December 1989, pp. 207–237.

language acquisition: Is it guided by a hypothesized system of Universal Grammar with certain particular properties? By a train of argumentation, it is hypothesized that if UG does guide adult SLA, then the Bounding theory principle of Subjacency ought to constrain the internal grammars of learners. By a further train of deduction, it is argued that if Subjacency does constrain the learner's grammar, then certain specific sentences (eg. *Who did you think that Mary saw?*) ought to be grammatical, while certain other sentences ought to be ungrammatical (eg. **Who did you believe the rumor that Mary saw?*)

A methodological question then arises: how is one to determine whether the learner's grammar in fact licenses a particular string of words? The purpose of this paper is to explore aspects of that question as it arises within the line of research just described. We will raise certain conceptual problems regarding the methodology which is typically employed in this research: that of obtaining from learners what are usually called "grammaticality judgements". We will then describe a methodology which can serve as a supplement to (though not a substitute for) such judgement data. This proposed methodology uses reaction time in a sentence matching task to probe the character of the learner's grammar, following the methodology of Freedman and Forster (1985). A small-scale study is then described which illustrates technique. The results suggest that the method can be useful, if it is carefully employed. Finally, we discuss the methodological problems in some detail and make concrete suggestions. Our focus throughout is on the methodology and logic of the study, not on the theoretical import of the results.

Some preliminaries. The research we are interested in here may be broadly characterized as UG-based. It should be borne in mind that the concerns we raise may not be equally applicable to other lines of research. We will use the term "sentence" to refer to a string of words whose status as a grammatical sentence is at issue. (A more scrupulously correct user of the term might balk at its use to refer to ungrammatical strings.) We will also use the common term "grammaticality" judgement to refer to the judgements speakers are asked to make, even though the meticulous user of technical language will correctly point out how misleading the term "grammaticality" can be (and will advocate other terminology, such as "acceptability" (Newmeyer, 1983, pp. 50-51)).

Reflections on Non-Native Judgement Data

We want to consider first some relatively high-level conceptual issues in the use of grammaticality judgements in SLA research, showing the urgent need for a supplement. It is not our purpose to review the the substantial body of SLA research which uses judgement data, and still less to discuss the full range of problems related to metalinguistic judgements in general. See Birdsong (1989) for an excellent review of these problems.

Grammaticality judgements as skilled performance.

A fortunate and rather surprising fact is that some (though not all) native speakers of a language often (though not in every case) have the ability to say, directly, whether a sentence is possible, ambiguous, possible under a certain interpretation, and so on. Linguists rightly take advantage of this ability when it is found. However, it must be emphasized that the existence of this ability is not an a priori necessary part of a linguistic system: It is quite easy to imagine systematic grammar-based communication systems which are very poorly designed for the task of making grammaticality judgements on arbitrary strings of words. In the case of human language (viewed as an idea-expressing and communicating system), it is hard even to see what particular (evolutionary?) functionality this ability might have. Thus its existence is might be considered a fortunate accident (fortunate for linguists). It is perhaps not surprising that not all native speakers seem to have developed this ability to the same degree. And this variation is by itself no reason to doubt the existence of grammatical systems.

Judgement giving seems in part to be a trainable skill. This is especially true in the case of determining whether a sentence is grammatical under a certain interpretation or in the closely related case of deciding whether a sentence is ambiguous. For example, beginning linguistics students often have difficulty seeing the ambiguity even of such completely uncontroversial examples as *The chickens are ready to eat*; yet by the end of the semester, such examples and others involving the interpretation of the grammatical functions of understood elements are easy for the students to evaluate. Perhaps they have learned what to look for. It would be absurd to say that their mental grammars had somehow changed during the the semester. On a personal note,

it took hours of hard work before one of us (R B-V) was able consistently to discern quantifier scope possibilities—for example, to discern that *Who did everybody see?* is ambiguous in a way that *Who saw everybody?* is not. Untrained naive informants are almost never able to give consistent judgements on these matters.

Judgement giving also seems to be degraded by various factors, even among the most skilled performers. For example, linguists sometimes report that it is hard to give reliable judgements if they are tired, if there are many distractions, and especially if they have been thinking intensively about a large number of closely related examples for some time.

In summary, judgement-giving is a skill, and judgements themselves are a kind of performance data. The ability to judge grammaticality is not a logically necessary part of grammatical competence, and even native speakers differ greatly in their skill in giving consistent judgements, especially in subtle cases.

The Assumption of Uniformity

Linguists who study the grammatical competence of native speakers are rightly not overly concerned with these problems. They rely on the judgements of the most skilled judges—who understand the task, who know what to look for, and who can make the judgements under the best circumstances. These ideal judges are often the linguists themselves. We need not poll naive speakers, who have unknown and highly variable ability to make judgements, in order to determine whether *The chickens are ready to eat* is ambiguous.

Native-language linguists then extrapolate from their own judgements and make conclusions about the mental grammars of the rest of the relevant linguistic community. In doing this they assume (at least initially, absent strong evidence to the contrary) that the mental grammars of others are not radically different from their own. There seems to be little reason to doubt the essential correctness and productivity of this procedure. (There have of course been some doubts expressed; usually these are based on misconceptions. See Newmeyer (1983, pp. 50–60) for relevant references and discussion.)

However, it is not often noted that the possibility of this extrapolation is dependent on an assumption about the nature of child language acquisition. We shall call this the assumption of the Uniformity of Resultant Systems (a phrase borrowed from Chomsky (1965, p. 58): the System Uniformity

Assumption for short). Speakers of a given language who have been exposed to roughly the same sorts of input during the development process all eventually come to possess essentially equivalent grammars. We say "essentially equivalent"—that is, equivalent in essence. Naturally, there can be legitimate dialect variation (there are double modals in Texas), and there can even be ideolectal variation on certain points (I say "davenport" and my sister says "sofa"). However, with regard to those aspects of "core grammar" which are of primary interest to scholars of Universal Grammar, one expects to find considerable uniformity. My sister's and my own grammars do not differ on whether *wh*-words can be extracted from relative clauses. One especially clear statement is that of Pinker, who proposes the following "external constraint" which any acquisition theory must meet: "A theory of a grammar for particular languages must be embedded in a theory of Universal Grammar that allows only one grammar … to be compatible with the sorts of sentences children hear." Thus, for example, when the linguist Robert May (1985, chapter 2) takes note of the quantifier scope facts alluded to above, he need not be overly concerned about that fact that Robert Bley-Vroman initially had trouble seeing the relevant contrasts. Even though Bley-Vroman may not be able consciously to perceive the contrasts at first (or ever), May can assume (if he cares) that Bley-Vroman's grammar is not really different. In language development both Bley-Vroman and May were exposed to the same sorts of sentences, and both underwent processes of language development guided by a Universal Grammar that allows only one grammar to be compatible with these sorts of sentences. Hence, both resultant grammars are the same with respect to the quantifier facts, even though one linguist may have an easier time seeing these facts than the other.

In the case of adult second language acquisition, the System Uniformity Assumption clearly does not hold. Learners of a given language who have been exposed to roughly the same sorts of input during the acquisition process cannot be assumed to have arrived at essentially equivalent grammars. The fact of wide inter-learner variation of a commonplace in second language acquisition research. (And it is a pillar of the Fundamental Difference Hypothesis. (Bley-Vroman, 1986)) The failure of the System Uniformity Assumption has the consequence that one cannot extrapolate from good judges making judgements under ideal conditions. One is forced to rely on poor

judges making judgements under bad conditions. This failure thus undermines the most important support for the use of grammaticality judgements in acquisition research.

The Assumption of Consistency

Furthermore, there are often cases where the judgements of the linguists themselves are uncertain. Linguists then will make deductions from the relatively clear cases to the less clear cases. Also, linguists will often be satisfied with one or two really good examples of some phenomenon, rather than insisting on a broad sampling of instances of the relevant structure. For example, a few clear cases of the impossibility of *wh*-movement out of relative clauses may be sufficient. If **What did the police arrest the men who were carrying?* is ungrammatical, so, probably is **What does your mother like a man who wears?* In making these moves, we assume that grammars are systematic objects, where *tout se tient*. Let us call this the "Consistency Assumption". For child-acquired native languages, it is probably reasonable to make this working assumption. In the case of adult-acquired foreign languages, there is considerable reason to doubt it. There is evidence that adult learners are much less consistent in their judgements than native speakers. Learners may often judge a given example of a structure as grammatical and then judge a slightly differently worded example of the same structure as ungrammatical. (Bley-Vroman, et al., 1988) Whether adult-acquired languages are systematic objects (in the same sense that child acquired languages are) is itself a controversial question in second language research. This controversy is most vigorously debated in precisely that area of second language acquisition research which is most likely to be interested in grammaticality.[1]

The Assumption of Dialect Separability.

A linguist, attempting to judge the grammaticality of an example in his/her native language, typically asks, introspectively, "Would/could I say this? How does this sound/feel to me?" Here the accent is on "I" and "me".

[1] A closely related assumption is also a crucial support for the use of grammaticality judgements: the assumption that the speaker's internal system is constrained by Universal Grammar. This means that it is possible to let the theory of Grammar make decisions in unclear cases. But, of course, in second language acquisition scholarship, whether the linguistic systems of adult learners are also constrained by UG is a controversial question—hence, not a reasonable working assumption.

That is, the concern is not with whether an example conforms to the norms of some external linguistic community , but with whether it conforms to the individual's internal grammar. The assumption here is that the judger can separate the question of whether "people say it" or of whether "it is correct" according to some norm from the essentially introspective question of whether it is grammatical according to the judge's own grammar. The assumption is that there is a way to reliably judge "my dialect" as opposed to "someone else's dialect" or "correct language." Let us call this the Assumption of Dialect Separability.

Is this a safe assumption to make in the case of non-native speakers? As researchers, we are of course interested in the character of the mental representation of grammar which underlies the learner's own use of the language. Yet learners also have ideas about whether a native speaker would or would not say something. Often these ideas are based on things they have read in books or learned in class, perhaps on their own sort of naive linguistic analysis of native speaker speech, or on folk wisdom. These ideas are not what we want to investigate. Put simply, we are interested in the learner's developing interlanguage grammar, not in the learner's conception of native speaker grammar. When the researcher asks the subject whether an example is grammatical, the learner may well interpret that question as asking whether the example is grammatical in the target language, not whether the example is part of the learner's interlanguage. Indeed, it may be too much to expect that learners can even separate such questions consistently. One might speculate that the failure of the assumption of dialect separability underlies many of the other difficulties in obtaining reliable judgements from non-native speakers. The lack of consistency in a given learner's judgements and the variation in judgements among learners may be in part due not so much to a lack of systematicity in a learner's developing internal grammar or to real differences among learners, as rather to the inability (and varying abilities) to separate the question of what is "grammatical for me" from what is grammatical in the target language.

A supplement to judgement data

Thus, several assumptions which are reasonable in native language research are unlikely to be safe in second language research. And these assumptions are precisely those upon which the use of grammaticality judgements depends. It is therefore prudent to seek additional means of probing learner's grammars. Naturally, all methods will have defects. However, different methods can be expected to be flawed in different ways, so that if results from various methods should converge on a conclusion, there will be correspondingly greater reason to believe that that conclusion is correct.

In our study, we explore the use of reaction time in a sentence matching task as one possible supplement to grammaticality judgements in SLA research.

Our experiment is based on the following phenomenon, first exploited in the psycholinguistic work of Freedman and Forster , whose work ours follows closely. (Freedman, et al., 1985) It is known, for a range of cases, that if a subject is asked whether two sentences match, the matching task takes longer if the sentences are ungrammatical than if they are grammatical. For example, if the example *The man saw the boy* is displayed on a computer screen, followed by the identical *The man saw the boy*, a subject will be able to say quite quickly that the two sentences are identical. If, on the other hand, the subject sees *Man the saw boy the* the followed by the identical *Man the saw boy the*, it will take significantly longer to determine that the sentences match.[2]

There have been two proposed explanations for this phenomenon. One plausible theory is that when the sentences are grammatical, the language processing system immediately and automatically produces a unified high-level representation of the examples; and identity can be determined on the basis of comparing unitary representations at this level. When the examples are not grammatical, however, no high-level representations can be computed, and matching must be done by other less efficient strategies, for example, by a word-by-word comparison. This idea has its roots in word matching tasks in

[2] In principle, there must be an analogous effect of grammaticality in determining that two non-matching sentences are in fact different. However, the speed of detecting a difference is also greatly effected by the locus of the difference in an example. If, for example, if the two sentences differ in their first words, the difference will always be detected very quickly, independent of grammaticality status.

reading research. Real words are faster to match than non-word letter sequences. This explanation is that advocated by Freeman and Forster (1985).

A second possible explanation is that when an example is ungrammatical, the mind constructs two representations: one of the example as is, another of a corrected version. In this second explanation it is mental correction rather than ungrammaticality per se which slows down matching time for ungrammatical examples. For a defense of this position, see Crain and Fodor (1987). Clearly, the debate over these two interpretations has important implications for the theory of language processing. For our purposes, the particular explanation will not be crucial. What will be crucial is the reality of the phenomenon.

Obviously, this phenomenon might be exploited to investigate grammaticality status independently from obtaining overt judgements, and might be used as a supplement to judgements. Indeed, Freedman, who did the pioneering research in this method, appears to have been motivated by a dissatisfaction with linguists' reliance on their own judgements as data.

The linguistic phenomenon

Our study explores an issue in learnability-theoretic markedness and SLA. (See White (1989) and Schachter (1989) for concise description of the basic concepts and their importance for the general theory of SLA.) The details are not important here; for we are concerned primarily with the methodological issues. For a more complete treatment of the linguistic and learnability theory within which our study is situated, as well as a more precise statement of the relevant linguistic facts, see Masterson (1988).

Korean differs from English in the sorts of NP structures which are permitted. English has a "functional category" of determiner. When a determiner is added to an NP, it "closes off category projection": in essence, once you've added a determiner to the NP, you're through—the NP is complete. As a consequence, multiple determiners are impossible (even where they would make sense: **This my sister's car is old, but that one is new*. (cf. *This car of my sister's is old, but that one is new*. Likewise, non-determiner modifiers cannot occur outside the determiner: **Red this car*. Korean does not distinguish a "functional category" of determiner as different from other modifiers in an NP. Korean modifiers which have a semantic determinative meaning (like the demonstratives, for example) do not close off category

projection. Hence, the ungrammatical English examples have grammatical Korean counterparts. The examples below illustrate the differences:

(a) *that Mary's sister [=that sister of Mary's]
*big John's car [=John's big car]
*Hard-working we [=we who are hard-working]
*John's Mary's criticism [=John's criticism of Mary]

(b) ku Mary-yu enni [that Mary's sister]
khu-n John-uy cha [big John's car]
bwuciren-ha-n wuli [hard-working we]
John-uy Mary-uy piphan [John's Mary's criticism]

The possibilities in Korean are a superset of those in English, which is "more restrictive." All the possibilities which exist in English also exist in Korean, and more besides. One possible theoretical account of the difference is that of Fukui (1986), who (in reference to an analogous contrast between Japanese and English) proposes that Universal Grammar provides a "Functional Categories Parameter". English is [+ functional categories] since it has determiner as a functional category. Korean is [–functional categories]: determiners are not a special functional category; they have the characteristic of ordinary modifiers. On learnability-theoretic grounds, the more restrictive setting of a parameter is the "unmarked", "initial hypothesis". English is "unmarked", Korean is "marked".

An important question in UG-based SLA research is whether learners will transfer a marked setting from their native language to their interlanguage during the process of foreign language learning. The prediction that learners will transfer a marked setting is a consequence of a general view of the place of Universal Grammar in foreign language learning. Specifically, will Korean learners of English assume that English is [-functional categories]? If they do, we deduce the consequence that their internal grammars ought to license English interlanguage structures like those illustrated above.

Experimental rationale and research hypotheses

We decided to explore the reaction time phenomenon described above in order to investigate the grammaticality status of the relevant structures in the learner's grammars. For this purpose, we constructed experimental example sentences of three types, as follows:

A. Grammaticality type #2: Grammatical English sentences (conforming to the [+ functional categories] setting of English) such as:

The cost of living is much higher in Japan.

NPs in such examples are also grammatical in Korean, in the relevant respect (since the Korean setting of the parameter is the less restrictive—"marked"—one). These examples are labelled as "grammaticality status #2" since they are grammatical in both (2) languages.

B. Grammaticality type #1. Examples which are ungrammatical based on a [+functional categories] grammar, but which would be grammatical in a Korean-type [- functional categories] grammar. These sentences included examples such as the following:

*John's Mary's photograph was hanging over the desk.

The mnemonic label "#1" indicates that these examples are only grammatical under one setting of the parameter—the Korean setting.

C Grammaticality type #0. Examples which would be grammatical under neither setting, for example:

Jane waited beside found the flower garden.

Table 1 shows how the types differ in grammaticality for the two settings of the functional categories parameter.

	[– FC]	[+FC]
#0	*	*
#1	ok	*
#2	ok	ok

Table 1: Consequences of differing settings of the Functional Categories Parameter: [±FC].

Note that a Korean-type [–FC] grammar would group types #1 and #2 together; while an English-type [+FC] grammar would group types #0 and type #1. We reasoned therefore, that if Korean speakers of English were carrying over their native language setting for the functional categories parameter into their English interlanguage, types #1 and #2 would be matched more rapidly than type #0 (and we might expect no significant difference in the matching times of types #1 and #2). On the other hand, if the Koreans set their interlanguage parameter at the English value, then types #1 and #0 would both be matched more slowly than type #2 (and there might be no significant difference in matching time between types #1 and #0). In this way, the reaction time patterning might shed light (from one angle at least) on an unknown and important aspect of the learners' developing competence.

For native speakers of English, of course, the grammaticality status of these examples is known. However, we also had a group of native English speakers in our experiment, as a standard against which to compare the Korean learners and as a check on the reliability of the experimental technique. In the English native speaker data, we reasoned that there ought to be a gap between the matching times for type #2 on the one hand and types #1 and #0 on the other.

Our design thus has a between-subjects factor of Native Language (Korean and English) and a within-subjects factor of Grammaticality Type (with three values: #0, #1, #2). The dependent measure is reaction time in the sentence matching task. We can then state the following three research

hypotheses:

1. Will there be a significant effect for Grammaticality Type? (There ought to be, or there is something radically wrong with the design.) Will the ungrammatical examples take the native speakers longer to match than the ungrammatical examples? (They ought to or, again, something is wrong.)
2. Will there be a significant effect for Native Language? (We expect there to be: Native speakers of English should be able to do the task faster than non-natives.)
3. Will there be a significant interaction of Native Language and Grammaticality Type? If Koreans are carrying over the native-language parameter setting, there should be a significant interaction: Type #1 would cluster with Type #2 for Koreans but with Type #0 for native English speakers. The latter two (ungrammatical) types ought to take significantly longer to match. On the other hand, if Koreans are adopting the English value for the parameter, then there should be no interaction of Native Language and Grammaticality Type.

For all statistical analyses, significance level is set at $\alpha = .05$.

Materials, subjects, and method of administration

We constructed a test instrument of 72 items, half of which were matching pairs and half of which did not match. In fact, we were only interested in the times for the matching pairs: the others were included to make the task realistic. The 36 items of interest (the matching pairs) were divided equally into the three grammaticality types, that is, there were 12 each of Types #1, #2, and #3. We attempted to construct examples of equal length and complexity and using simple vocabulary. In particular, we were very careful not to make ungrammatical examples longer or more complex than the corresponding grammatical examples. In many cases, this was quite easy. For example, Type #2 *that green car* can be converted to Type #1 *green that car* simply by inverting words. (We return to the problems of controlling for length and complexity below.)

Our subjects were 14 native speakers of Korean, participating in the Pacific-Asian Management Institute (PAMI), a summer program at the

University of Hawai'i[3] for international business leaders.

Our native English speaking group included University of Hawai'i students, professors, and others we were able to recruit individually. All were adults. There were 14 native speakers in all.

The experiment was administered on personal computers[4] (MS-DOS machines with 12-inch monochrome monitors). First, a set of instructions and example sentences was displayed, then a trial set of 5 items, then the actual test instrument. A given test item is presented as follows:

1. The first of the pair of sentences is displayed in the upper center of the screen, with ordinary capitalization (not all caps).
2. After a short delay (see below), the second of the pair appears about one inch beneath it, and offset about one inch to the right. (The offset is to prevent visual vertical matching of word-shapes.) Now both sentences are visible on the screen.
3. If the sentences match, the subject is to press the "J" key on the keyboard; if they do not match, the "F" key is to be pressed.[5] The "J" key has a blue dot pasted on it, and the "F" key has a red dot pasted on it. At the top of the screen the reminder "RED = different BLUE= identical" is displayed at all times.
4. After the key is pressed, the subject's response is echoed at the bottom of the screen. For example, if the subject has pressed the red key, the words "RED DIFFERENT" appear at the bottom left of the screen. If the subject presses any key except the two designated keys, the computer "beeps", a reminder is displayed, and the subject must try again until either the "J" or "F" key is pressed.
5. At the bottom of the screen, the words "Press SPACE-BAR for next item"

[3] We are grateful to Martha Pennington and to Jeong-Ryeol "Jay" Kim for helping us administer the experiment, to Gerald Chang for his help in arranging for the personal computers, and especially to the Korean subjects for their cheerful participation. We gave a party for the participants after the experiment: everybody had a good time.

[4] For information on obtaining the Pascal source code, contact Robert Bley-Vroman, who wrote the program.

[5] No attempt was made to set things up differently for left-handed and right-handed subjects. Many subjects in SLA experiments come from cultures in which left-handedness is strongly stigmatized, so it is difficult to get accurate information about the preferred hand. We felt the possible benefits of adjusting for handedness were not worth the cross-cultural hassle.

now appear. The subject presses the space bar when ready—the subject can rest as long as desired—and the next item appears (see step 1).

At the beginning of the test, we inserted a single "fake" item, which was not part of the actual test. We had noticed during early trials that the first item in the test always took much longer than all subsequent items. Giving an initial filler item corrected this problem. The same initial filler was used for all administrations.

The actual items were displayed in random order—a new order was generated by the computer for each subject.

For every item, we recorded (a) the response latency in milliseconds, measured from the time the second sentence of the pair was displayed until one of the designated keys was pressed; (b) the serial order in which the item had been presented on that experimental run; (c) whether the response was correct.

Analysis, results, and discussion

For each subject we computed a mean response time for each of the three Grammaticality Types. In accordance with accepted practice for this type of measure (See Freedman, et al. (1985, p. 110).), we adjusted values which were more than two standard deviations from a given subject's mean for a given type back to the two-sd value. (This technique is important in this type of response latency measurement, since a subject can occasionally get distracted and for some reason fail to press a key for a very long time—we sometimes got times of a half a minute or more for an item. However, in practice it was only very rarely necessary to adjust a measurement in this way.) We ignored all cases in which the subject responded incorrectly or pressed some key besides the "F" or "J" keys. In fact, this meant throwing out less than 0.5% of the trials.

Tables 2 and 3 show the results for native English-speaking and Korean subjects respectively. Times given are the mean response latencies (in milliseconds) for the subjects. Standard deviations among subjects are also displayed.

Type	Mean Response Time (msec)	Standard Deviation
0	1737	365
1	1783	364
2	1633	369
0 vs. 1 p>.05; 0 vs. 2 p<.05; 1 vs. 2 p<.05 (Scheffe F; repeated measures ANOVA)		

Table 2: Native speakers of English
Response times for 3 Grammaticality Types.

Type	Mean Response Time (msec)	Standard Deviation
0	4146	1248
1	4283	1047
2	3904	1140
0 vs. 1 p>.05; 0 vs. 2 p>.05; 1 vs. 2 p<.05 (Scheffe F; repeated measures ANOVA)		

Table 3: Native speakers of Korean
Response times for 3 Grammaticality Types.

As expected, there is a significant effect for Native Language. The native speakers respond much more rapidly. There is also a significant overall effect for Grammaticality Type.

Within the native speakers, we also see a very clear effect for Grammaticality Type. This effect is as predicted in the experimental rationale. Compare Table 2 with the right-hand side of Table 1. In English, it is known that Type #2 represents grammatical examples, while Types #0 and #1 are both ungrammatical. Correspondingly, in planned comparisons, we see that Types #0 and #1 each take significantly longer to match than Type #2 (and there is no significant difference between Types #0 and #1). The English setting of the Functional Categories parameter is thus being reflected precisely as predicted in the response times in the sentence matching task.

There is no significant interaction of Native Language and Grammaticality Type. Both Koreans and native English-speakers show the same pattern. The Korean subjects, like the native English speakers, take longer on Types #0 and #1 than on Type #2. (However, in the case of the Koreans, the difference between #0 and #2, though in the predicted direction, is not statistically significant.) The Korean subjects show a pattern of response latencies which suggests an English-like setting of the Functional Categories parameter.

Our ANOVA on subjects leads us to believe that the results will generalize; that is, the observed differences are not just a result of the chance choice of subjects. However, a major limitation of our study is that we are not certain that the results will generalize to all items of a given type; that is, it is possible that the observed differences were in part the result of our choice of items. (While ANOVA on subjects yields statistically significance; ANOVA on items does not.) We return below to the problem of how to overcome this limitation. Basically, we believe that the solution lies in much more careful item construction and increased sophistication of statistical design and analysis.

General Discussion

The substantive issue of the acquisition of the functional categories parameter is not the focus of this paper. Still, it appears from our data that the English-like setting of the parameter can be achieved by native speakers of a [–Functional Categories] language. This result is of some interest since in certain current theories of foreign language learning such a parametric resetting from unmarked to marked ought not to be possible, or ought not to be possible without some kind of negative evidence or explicit instruction (White, 1989; Bley-Vroman, 1986). Naturally, we need more information about the subjects (including especially "proficiency" level and language learning history); we need experiments with different subjects and different linguistic backgrounds[6], and especially experiments of longitudinal or cross-sectional design.

[6] It would also be valuable, obviously, to have data of native speakers of Korean matching Korean examples. At the time this experiment was done, we had not yet perfected a technique of displaying Hangul script on the screen during the experiment. This obstacle has now been overcome. We will now also be able to test native speakers of English learning Korean.

In an earlier unpublished pilot study by one of us (Masterson), grammaticality judgements had been obtained from Korean learners on examples relevant to the functional categories parameter.[7] The results of that study had suggested that the Koreans' interlanguage grammars of English might treat Type #1 sentences as ungrammatical (i.e. that they had successfully "reset the functional categories parameter"). We had not had full confidence in those results in part because of the concerns about grammaticality judgements of non-natives which we have expressed here (also, because of certain design problems in the study). The fact that the reaction time results correspond to the grammaticality judgement results provides support for both. Again, we reason that although both methods are flawed, they are clearly flawed in different ways, so that the convergent results inspire a degree of confidence which neither set of results commands by itself. The convergent results also give confidence in the validity of the methodology of reaction times, which is the major focus of research.

We now consider several aspects of that methodology which arose during the design, running, and analysis of the experiment.

The problem of item construction: controlling for length and complexity

Processing time is affected by many things; grammaticality is only one of them: length is another—also various sorts of syntactic complexity, semantic plausibility, familiarity of vocabulary, etc. In constructing items for an experiment it is important to ensure that, for example, items of a particular grammaticality type are not all longer or shorter than those of a different type. In our initial experiment, we approached this problem unsystematically. We merely attempted to avoid large and obvious differences. Therefore, we think, but we are not really certain, that the observed differences in reaction time are the result of grammaticality, and not of some accidental other property which distinguishes the types.

Clearly, more systematic techniques of example construction must be devised. The approach which is appropriate will depend in part on the kind of structural property being studied. In the case of the possible orders of elements, such as the ordering of determiners and adjectives in NPs, it is

[7] The study was originally conceived at a workshop on parameter setting and language acquisition at the 1987 LSA Institute at Stanford University.

possible to create corresponding grammatical and ungrammatical pairs of precisely the same length and complexity (and vocabulary), merely by reversing words. For example:

(a) Mary saw that old man.
(b) Mary saw old that man.

Many things which SLA researchers want to study are of this sort. White (1989), for example, studied the grammaticality status of examples like the following pairs in English learners of French and French learners of English:

(a) Jane often eats ice cream.
(b) * Jane eats often ice cream.

Lynn Eubank (1989) has been exploring the use of reaction times to study the grammaticality status of examples like the following in the German of English-speaking learners. In these examples, too, grammaticality status is a function of an ordering difference:

(a) Der Mann zieht seine Jacke am Nachmittag an.
(b) * Der Mann zieht an seine Jacke am Nachmittag.

While it is relatively easy to create corresponding grammatical and ungrammatical items in the case of simple ordering effects, other properties are much more difficult to work with. For example, the possibility of null objects in languages like Chinese and Japanese might lead a researcher to be interested in Chinese/Japanese interlanguage status of English examples like the following:

(a) John devoured the pancakes.
(b) * John devoured.
(c) John put the cake on the table.
(d) * John put the cake.

Here the ungrammatical example of the pair is created by deleting words from the grammatical example. The differing lengths of the items may be

expected to have an effect on processing time. There are at least two approaches to dealing with such difficulties. The first is to lengthen the ungrammatical examples by adding material. For example, instead of **John put the cake* one might use **John put the cake in the evening*. The ungrammatical example is now of the same length as the grammatical example; also, the additional material is a prepositional phrase (as in the grammatical example). However, the equivalence of the grammatical and ungrammatical examples is only superficial. Since *in the evening* is an adjunct rather than a complement to *put*, it should probably be assigned a different constituent structure. Furthermore, if the ungrammatical example were in fact being treated as grammatical by Chinese learners because of transfer of Chinese grammar, the learners might be assigning it a structure like the following, where the first Ø represents a null topic:[8]

[Ø [John [put the cake Ø] in the evening]]]

On the other hand, *John put the cake on the table* might be assigned a (simpler) structure, like this:

[John [put the cake on the table]]

Suppose the Chinese learner of English had an English interlanguage grammar which licensed both *John put the cake on the table* and *John put the cake in the evening*. But that grammar might assign a more complex structure to *John put the cake in the evening*. This additional complexity might lengthen reaction time and thus give the impression (false, in this case) that *John put the cake in the evening* was ungrammatical according to the learner's grammar.[9]

One might also try to create equivalent examples by changing verbs. Thus, one might compare examples like the following:

(a) John ate.
(b) * John devoured.

[8] We follow here, in spirit, the proposal that the missing complements of verbs are actually bound by a null topic (Huang, 1984).

[9] Also, the special pragmatic conditions which are required to make null arguments interpretable might slow down processing.

However, there is now a danger that some irrelevant feature of the lexical change might contribute to processing time. For example, the subjects might be less familiar with the word *devour* than with *eat*.

The general second strategy to dealing with inevitable differences in length and complexity is to try to factor out the differences, leaving only the difference of interest. For example, suppose one is interested in the difference between *John liked the ice cream* and **John liked*. One could set up a control pair of items which differed in length, but not in grammaticality. For example, one could use:

(a) John ate the ice cream.
(b) John ate.

(c) John liked the ice cream.
(d) * John liked.

One then compares the difference between (a) and (b) with the difference between (c) and (d). The (a) vs. (b) difference shows the effect of length alone, while the (b) vs. (c) difference shows the effect of length plus grammaticality. But subtracting the length difference, one arrives at an estimate of the effect of grammaticality alone.[10] An analogous technique could be used to control for any difference caused by vocabulary substitution. For instance, a comparison of *John devoured the pancakes* and *John ate the pancakes* could give a value for the effect of the word substitution, independent of grammaticality.

Naturally, it would be possible to combine techniques. One can add material to sentences in order to attempt to balance length and complexity and also attempt to control for any inadvertent changes caused by the addition alone. In this way designs might become quite complex indeed.

Note finally that the problems of correcting for length and complexity are more severe in the case of grammaticality contrasts which involve substantial abstract difference in structure. For example, consider the problem of comparing extraction of wh-words from relative clauses versus clausal

[10] There are various statistical approaches to this correction. Analysis of covariance might be appropriate. Crain and Fodor simply calculate an "increment" by subtraction (Crain & Fodor, 1987, especially Table 1 and Table 4).

complement to verbs (*Who did you say John likes* vs. **Who do you admire the girl who John likes*), or of extraction from subject position vs. extraposed subject position (*Who did it surprise him that you like* vs. **Who* (*did*)*that you like surprise him*). Here it is extremely difficult to control for all relevant factors.[11]

An advantage for SLA researchers. In general, excessive complexity in experimental design is to be avoided, and it is possible that some of the problems in creating corresponding pairs of differing grammaticality status may be less severe for SLA researchers than for students on native language processing. Suppose one is interested in the difference between *John devoured the pancakes* and *John devoured*. To be sure, one is longer than the other. In fact, because one is long but grammatical and the other is short but ungrammatical, they might not be significantly different in processing time (indeed, in the worse case, the ungrammatical one might even take less time). However, as SLA researchers our goal is not to determine the grammaticality status of these examples in native English (we know that) nor to determine how much of the reaction time of these two pairs is due to grammaticality status. Rather, we are typically interested in whether a non-native speaker's grammar differs from the native speaker's grammar. Determining this is a simpler task. Our technique is to compare the native speaker latency data with the non-native data. In the hypothetical case, suppose that for native speakers *John devoured* even takes a bit less time to process than *John devoured the cake* take (the ungrammaticality being out-balanced by shortness). If a non-native speaker takes A LOT less time to process *John devoured* than *John devoured the cake*, this may mean that *John devoured* is not only shorter but also grammatical. (There are various statistical approaches one might take here. The simplest would be to look for a statistically significant interaction of grammaticality type and native language, as was done in our experiment.)

In recommending this approach, we wish to emphasize that the best course is to control for length and complexity directly, by making sure that grammaticality contrasting examples are equivalent in all other relevant respects. In many cases of interest, especially those simply involving serial order, this is not difficult to do. All other approaches must make additional

[11] Variability introduced by length and complexity is even a problem if it is evenly distributed among grammatical and ungrammatical sentences: it creates bigger standard deviations and makes it more difficult to get statistically significant results with a small n-size.

assumptions, which may or may not be justified[12]; and they add the to complexity of the experimental design. The more complicated the design, the more things there are to go wrong.

Preventing inadvertent content-based processing differences

It is natural for an experimenter to want the examples to sound plausible. In the grammaticality tasks of Bley-Vroman, et al. (1988) there are examples like **What did John say that would fall on the floor, if we're not careful*; **What sort of food is to digest easy*; *Which bed does John like to sleep in*. Examples with abundant sentence-internal are constructed, so that, if the sentences are indeed grammatical for the subjects, the subjects will accept them, and not reject them because they are a little hard to interpret. Thus, *…if we're not careful* is supposed to provide contextual support for the intended interpretation of the *fall-on-the-floor* example. By putting *bed* together with *sleep* and *digest* with *food*, the examples gain coherence. One would not want to use an example like *What kind of rock is to digest easy*. Indeed, in grammaticality judgement tasks, the tendency has been to use highly contextualized examples rather than more "neutral" examples—to use *What kind of food is easy to eat* rather than, say, *What kind of thing is easy to eat*. The unstated rationale is probably that the burden of proof should fall on ungrammaticality. Very likely, this philosophy stems from a common experience of linguists working on their own native language: A structural configuration may initially seem ungrammatical, but if the correct context is given, and if an example is carefully worded, then the grammatical possibility becomes clear. (See Birdsong (1989, passim, and esp. pp. 64, 68) for discussion of context effects on judgements.) Let us grant that this rationale makes some sense for grammaticality judgements. It is probably not equally valid for response latency experiments. No doubt, contextual plausibility effects processing. It is unlikely that the researcher will be able to make all examples equally plausible. The probably inevitable variation in plausibility among the examples will introduce an unknown amount of variability in the response times. It may result in such large variance that real differences of interest may not be detected, having been swamped by irrelevant

[12] For example, the idea that the comparison of natives and non-natives can effectively control for the effects of length assumes that those effects are independent of native language background. This seems reasonable, but it is not necessarily correct.

variation.[13] Very large numbers of examples and/or subjects may then be necessary.

It is therefore probably wise to choose relatively neutral examples rather than to go to some effort to create plausible examples. One may wish to consider constructing all examples from some small set of rather colorless vocabulary. If this is done, it may even be possible to generate examples by mechanical algorithm, and thus avoid other forms of inadvertent experimenter bias. It would have the added benefit of alleviating concerns about the generalizability of the results beyond the particular examples used in the study.

Technical issues in administration of the experiment

The filler item. As discussed above, we inserted a filler item at the beginning of the test to get the subjects started. We recommend that this always be done. In our experience, a single item is sufficient; however other scholars have used more. In one experiment, Crain and Fodor placed 8 filler items at the beginning of every set (Crain, & Fodor, 1987).

Number of items. Our study demonstrates the possibility of getting some results from as few as 12 experimental items of each grammaticality type. However, this number is probably too small. The corresponding native-language psycholinguistic research has typically used about 20–30 items per type. A reasonably large number should be aimed for because of the considerable variation in times among items of a single type. With very careful item construction, some of this variability might be eliminated, but some will always remain. The number of items also relates to the generalizablity of the results—a point to which we return below.

In our study, the total number of items (both experimental items and distractors) to which each subject responded was 72. This too is in line with other work, where the number of items has typically ranged from about 60 to 90. Most subjects will finish the experiment in less than a quarter hour. We

[13] Of course, there is also the obvious error of making the grammatical sentences more plausible than the ungrammatical ones, or vice versa. If some clear, relatively mechanical procedure is adopted to produce the pairs, as would be anyway done to avoid length/complexity problems, this error can be avoided.

fear that subjects may become fatigued if more items are used. It is possible that for non-native speakers a smaller number would be more reasonable. In the analysis of our results, we found no significant correlation between reaction time and order of presentation; items presented at the end did not take longer to match than earlier items. At least, if there was fatigue, it did not reveal itself in a general slowing down (or speeding up).[14]

Ratio of types of items. In our study the ratio of matching examples to non-matching examples was 1:1. Half of the items were matching pairs; half were non-matches. This balance is intended to prevent any response bias which might develop if there were unequal numbers. Alas, only the matching items are of interest to the experiment; the others are really distractors included simply to make the task realistic. It would certainly be desirable to reduce the number of non-matches if possible. With fewer non-matches, one could include more experimental sentences without creating an unduly long test. Crain and Fodor have used fewer non-matching items than matching items—a ratio of 3:1—and they report no ill effects (Crain, et al., 1987). Upon reflection, it is clear that even if there is a response bias toward matching, such a bias is really no problem unless the grammaticality effect itself is somehow compromised. Studies should be done to discover how small the number of non-matches can safely be. There is as yet no compelling evidence that a 1:1 ratio must be used.

Number of subjects. We had 14 subjects each for the two language groups (28 subjects total). While this is broadly in line with the number of subjects in comparable native language groups, we feel it probably represents a minimum.[15] The statistical significance of the differences of the sort we

[14] In his research, Eubank has found a relationship between reaction time and order of the items on the test: subjects were faster at the end (Eubank, 1989). If the relationship between reaction time and presentation order were simple (linear, say) and independent of grammaticality status, then one might try to make a statistical correction for the effect of presentation order. It is also possible that fatigue would show itself in the ability to distinguish between grammatical and ungrammatical examples, rather than (only) in task speed. The sensitivity of the test to grammaticality status would then partly depend on the position in the test of the examples. The test might become less (or more) sensitive toward the end. In an ANOVA design, one could look for an interaction of grammaticality type and presentation order.

observed can just barely be detected with this number of subjects. One of our results (one which is consistent with our interpretation of the results) is the reaction time difference between types #0 and #2 for Koreans (see Table 3); but this difference was not statistically significant. Had we had about twice as many subjects, it would have been significant (assuming, of course, that the additional subjects would behave the same). We suggest about 30 subjects per native language group in an experiment of this type would be reasonable; that is, a given experimental sentence should be responded to by at least 30 subjects.

Timing of item display. For non-native speakers, we held the first of a pair of examples on the screen for 4 seconds before the second item appeared; for native speakers we waited 2 seconds. It seems to us reasonable to choose a display time which will allow the most subjects to read the item comfortably. Nonetheless, we do not want to make it too long, lest some subjects go back and study the example, or have their minds wander from the task. We chose our times by informal piloting. A native speaker would do the task and tell use whether it seemed rushed, too slow, about right; the same with non-natives. We made a couple of adjustments till we came up with something which seemed about right. A much more systematic way of deciding exposure time would be desirable. One idea (which we are now exploring) is to run an initial calibration study in which subjects themselves control the appearance of the second item. We would then use that data to decide on a reasonable uniform display time for the experiment. (One criterion would be to set it at 1 standard deviation above the observed mean time on the calibration run.[16]) No doubt, native and non-native display times should be determined independently.

In our study, after the second item of the pair appears, both the items are held on the screen together until the subject responds. An alternative is to

[15] Freedman and Forster used groups of 15 subjects. (The experimental sentences were divided up among four groups, so that the total subject pool required was 60.) In Crain and Fodor's experiment 2, each version of each experimental sentence was responded to by 12 subjects (Crain and Fodor, 1987, fn. 20, p. 145).

[16] If there is very great variability among subjects on this calibration run, it would make one worry about the possible effect (uncontrolled) of intersubject differences in reading speed. It would of course be possible to allow subjects to control the time of appearance of the second of the pair even in the actual experimental run. We know of no experiments which have done this, and its effect is unknown.

blank the screen after some designated time, whether or not the subject has responded. The rationale for this technique would be to put pressure on the subjects and to discourage them from employing non-syntactic matching techniques (for example, trying to compare left-to-right, word-for-word). In fact, Freedman and Forster use this alternative, blanking the screen after both of the paired items have been displayed for two seconds. We now regret our decision.[17]

Feedback. We did not give feedback on whether the response was correct. While it is not clear what effect feedback has, Crain and Fodor note that their obtained reaction times, using a design like ours which did not use feedback, were much longer on the average than those of comparable experiments by Freedman and Forster which did use feedback (Crain, et al., 1987, p. 145). Apparently, subjects can perform the task faster with feedback. We do not know whether feedback will enhance or diminish the grammaticality effect (or whether it even matters at all). All else being equal, one ought to give feedback if it makes the experiment faster and easier for the subjects.

Analytical and design issues

In our study, we had three types of sentences and each subject responded to all three types. An alternative approach is to have different subjects respond to different types. There are advantages and disadvantages to each of these two approaches. The choice between them involves interesting trade-offs.

If a given subject responds to all grammaticality types, there is more information in the results than if different subjects respond to different types. There is, typically, great variability in overall speed among subjects. Some will have average speed of less than a second on the task; others will take 4 seconds. If different subjects respond to different types, the great intersubject variability within each group may overwhelm the relatively small differences among the groups. If, on the other hand, each subject responds to each type, one can see even a relatively small difference between types, to the extent that each subject shows it consistently. In our analysis we used repeated measures ANOVA to take advantage of the fact that a given subject is "repeatedly

[17] Eubank (personal communication), who has tried out both techniques, believes the grammaticality effects are clearer if the screen is blanked.

measured" on reaction time under different conditions: for each subject we had three figures.[18] If we had not had the additional information which a repeated measures design gives, none of our results would have been statistically significant. Put differently, without a repeated measure design, we would have had to have had far more subjects to see clear results. The classical advantages to repeated measures designs are "increased precision and economy of subjects" (Stevens, 1986, p. 405).

On the other hand, there is a clear drawback to a repeated-measures design for an experiment of the sort we are considering. If each subject is to be exposed to every type of sentence, then the number of items of each type must be kept low or the test will get too long. Because we were forced to use relatively few examples of each type, and because there may be considerable variability among the items of a given type, we may not be able to demonstrate statistically that our results generalize beyond these particular items (that is, we can show a statistically significant effect for grammaticality for subjects, but not for items). On the other hand, when separate groups respond to different types, many more examples of each type can be included.

The advantage of the repeated-measures design is its precision and its economy of subjects. The advantage of independent subject samples for each condition is the possibility of using many examples of each type. If it is difficult to obtain large numbers of subjects, or if intersubject variability is likely to be great, then one would lean toward a repeated-measures design. If, however, it is felt that many items will be necessary for a given type, grammaticality types might have to be divided up among groups.[19]

In second language acquisition research, subject groups will often have to be small and there will usually be a great deal of intersubject variability: there always is, among language learners. Thus, it will make sense to have each subject respond to all grammaticality types. It will be correspondingly especially important to "make every item count." Items must be very carefully constructed to avoid all sources of noisy variation. Tests should be carefully piloted. Of special concern should be pilot runs with native speakers. If one

[18] The even simpler technique of a paired-samples t-test would have been appropriate if we had had only two grammaticality types.

[19] The researcher should also consider more elaborate possibilities. One might have one group respond to types #1 and #2; another to #2 and #3; and a third to #3 and #1. The two alternatives discussed here are the ends of a continuum of possibilities.

cannot get good, clear results with native speakers, where the state of the internal grammar is presumably known, one cannot expect to be able to say anything with confidence about non-natives.

Generalizablity and the need for a composite F. One should do analysis of variance both on items and on subjects. Only then might one safely conclude that results are (1) not limited to the particular subjects, and (2) not limited to the particular items. In our case, ANOVA on subjects yielded statistically significant results, but ANOVA on items did not. Technically, even separate ANOVAs on subjects and items is not sufficient. Ideally, a composite F should be calculated, simultaneously taking subjects and items into account, following the method of (Clark, 1973). In fact, in SLA research, ANOVA is almost always restricted to subjects alone, and, to our knowledge, Clark's composite F has never been used.[20] If, in designing future studies, one adopts the general approach advocated here, Clark's composite F would be quite reasonable. Our speculation is that if items are designed with great care, if there are sufficient numbers of items and of subjects, and if the experiment is planned and administered with great attention to detail, it will be possible to generalize both from the experimental subjects to the relevant population and from the experimental items to all items of the relevant type.

Received December 22, 1989

Address for correspondence:

Robert Bley-Vroman
Department of ESL
University of Hawai'i
1890 East-West Road
Honolulu, HI 96822

[20] This failure of statistical analysis, which Clark sees as a kind of "fixed-effects fallacy", afflicts most acquisition research (not just SLA research) which attempts to compare subject behavior on different kinds of structures.

References

Birdsong, D. (1989). *Metalinguistic Performance and Interlinguistic Competence*. Berlin: Springer-Verlag.

Bley-Vroman, R. (1986). The logical problem of foreign language learning. University of Texas ms. [To appear in *Linguistic Analysis*].

Bley-Vroman, R., Felix, S., & Ioup, G. (1988). The accessibility of Universal Grammar in adult language learning. *Second Language Research, 4*(1), 1–32.

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

Clark, H. 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.

Crain, S., & Fodor, J. D. (1987). Sentence matching and overgeneration. *Cognition*, *26*, 123–189.

Eubank, Lynn. (1989). Sentence matching, German, and Universal Grammar. Paper presented at the Boston Univerisity Conference on Language Development, Boston, MA, October 15, 1989.

Freedman, S., & Forster, K. I. (1985). The psychological status of overgenerated sentences. *Cognition*, *19*, 101–131.

Fukui, N. (196). *A Theory of Category Projection*. PhD dissertation. MIT.

Huang, C.-T. J. (1984). On the distribution and reference of empty pronouns. *Linguistic Inquiry*, *15*(4), 531–574.

Masterson, D. (1988). Functional categories and parameter setting in L2 acquisition. *University of Hawai'i Working Papers in Linguistics*, *20*(1), 1–12.

May, R. (1985). *Logical Form: Its Structure and Derivation*. Cambridge, MA: MIT Press.

Newmeyer, F. S. (1983). *Grammatical Theory: Its Limits and Possibilities*. Chicago: University of Chicago Press.

Schachter, J. (1989). Testing a proposed universal. In S. Gass, & J. Schachter (Ed.), *Linguistic Perspectives on Second Language Acquisition* (pp. 73–88). Cambridge: Cambridge University Press.

Stevens, J. (1986). *Applied Multivariate Statistics for the Social Sciences*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

White, L. (1989). The Adjacency Condition on Case assignment: Do L2 learners observe the Subset Principle. In S. Gass, & J. Schachter (Ed.), *Linguistic Perspectives on Second Language Acquisition* (pp. 134–158). Cambridge: Cambridge University Press.