# THEORETICAL AND METHODOLOGICAL PERSPECTIVES ON THE USE OF GRAMMATICALITY JUDGMENT TASKS IN LINGUISTIC THEORY

ANNIE TREMBLAY

*University of Hawai'i at Mānoa*

The aims of present paper are: (a) to examine the theoretical and methodological issues pertaining to the use of grammaticality judgment tasks in linguistic theory; (b) to design and administer a grammaticality judgment task that is not characterized by the problems commonly associated with such tasks; (c) to introduce FACETS as a novel way to analyze grammaticality judgments in order to determine (i) which participants should be excluded from the analyses, (ii) which test items should be revised, and (iii) whether the grammaticality judgments are internally consistent. First, the paper discusses the concept of grammaticality and addresses validity issues pertaining to the use of grammaticality judgment tasks in linguistic theory. Second, it tackles methodological issues concerning the creation of test items, the specification of procedures, and the analysis and interpretation of the results. A grammaticality judgment task is then administered to 20 native speakers of American English, and FACETS is introduced as a means to analyze the judgments and assess their internal consistency. The results reveal a general tendency on the part of the participants to judge both grammatical and ungrammatical items as grammatical. The FACETS analysis indicates that the grammaticality judgments of (at least) two participants are not internally consistent. It also shows that two of the test items received from six to eight unexpected judgments. Despite these results, the analysis also indicates that overall, the grammaticality judgments obtained on each sentence type and on grammatical versus ungrammatical items were internally consistent. In light of the results and of the efficiency of the program, the implementation of FACETS is recommended in the analysis of grammaticality judgments in linguistic theory.

## INTRODUCTION

Grammaticality judgment (GJ) tasks are one of the most widespread data-collection methods that linguists use to test their theoretical claims. In these tasks, speakers of a language are presented with a set of linguistic stimuli to which they must react. The elicited responses are usually in the form of assessments, wherein speakers determine whether and/or the extent to which a particular stimulus is "correct" in a given language. The use of GJ tasks in linguistic theory is necessary because it provides a means to: (a) assess the speakers' reactions to sentence types that only occur rarely in spontaneous speech; (b) obtain negative evidence on strings of words that are not part of the language; (c) distinguish production problems (e.g., slips, unfinished utterances, etc.) from grammatical production; and (d) isolate the structural properties of the language that are of interest by minimizing the influence of the communicative and representational functions of the language (Schütze, 1996). Crucially, neither elicited production tasks nor naturalistic data collection provide a means to accomplish (a)–(d).

In recent years, however, the reliance of linguists on GJs to formulate, support, or refine their theoretical claims has become quite controversial. One concern which has emerged in the literature is the absence of clear criteria to determine what is the exact nature of grammaticality. This has led to a general division in the field between linguists who treat grammaticality as a dichotomous concept and those who treat it as a gradient concept. Another concern has to do with the validity of GJs, that is, the extent to which they actually reflect grammatical competence. This concern arises because: (a) GJs, just like other kinds of data gathering tools, are influenced by extragrammatical (or performance) factors, and (b) it is not clear what the relationship is between objectification skills, such as metalinguistic judgments, and what is being objectified, here grammatical knowledge. Last but not least, the use of GJ tasks has been contested, because such tasks too often lack the rigorous control techniques normally used in psychological experimentation. These include, among others, providing maximal control for the effect of extragrammatical factors and assessing the reliability of GJs.

The present paper examines the above theoretical and methodological issues from a critical perspective in order to design and administer a GJ task that is not characterized by the problems commonly associated with such tasks, and it proposes a novel way to assess the internal consistency of GJs. The paper situates GJs within linguistic theory, first by summarizing the different positions adopted on the nature of grammaticality, and second by discussing important validity issues pertaining to the use of GJ tasks in linguistic experimentation. Third, it provides a description of the measures to be taken in the implementation of methodologically sound GJ tasks; the discussion tackles the creation of materials, the procedures of the task, and the analysis and interpretation of the results. Following these recommendations, a GJ task is designed and administered to a group of native speakers of American English. The program FACETS is introduced as a means to analyze these (and other) GJs in order to determine (a) which participants should be excluded from the analyses, (b) which test items should be revised, and (c) whether the GJs are internally consistent. This novel method of analysis will prove essential to solving some of the methodological problems commonly associated with GJ tasks and determining whether any firm conclusions can be drawn from the study. Finally, the paper provides some directions for future research in the design and implementation of GJ tasks. Given the scope of the issue at hand, the discussion is limited to the elicitation of GJs from native speakers.

## THEORETICAL ISSUES

### *Grammaticality: Dichotomous or Gradient?*

The absence of clear criteria to determine what is the exact nature of grammaticality has raised a few concerns on the part of some researchers (e.g., Schütze, 1996). The assumption underlying most formal models of grammar is that grammaticality is *dichotomous*, with structures being categorized as either grammatical or ungrammatical. For example, in the Minimalist Program (Chomsky, 1995), sentences converge if they follow the wellformedness rules of the language, and crash if they don't. Such a model is largely incompatible with the notion of grammaticality as a *gradient* concept. Chomsky (1975) himself, however, acknowledges that grammaticality can come in varying degrees: "There is little doubt that speakers can fairly consistently order new utterances, never previously heard, with respect to their degree of 'belongingness' to the language" (p. 132). That linguistic structures are not always equally (un)grammatical is in fact reflected in linguists' own annotations (e.g., [?], ?, ??, *, **), as shown in (1a)–(1d):

(1) a. Not to be invited is sad.

b. (?)To have not had a happy childhood is a prerequisite for writing novels.

c. ?To be not invited is sad.

d. * To get not arrested under such circumstances is a miracle.

(Haegemann & Guéron, 1999, p. 308)

The use of multiple (un)grammaticality annotations, albeit inconsistent, is motivated by the necessity to represent intermediate levels of (un)grammaticality originating from two sources: *interspeaker* variation and *intraspeaker* variation. On the one hand, speakers sharing different idiolects tend to vary from one another with respect to their judgments on particular sentences. Accordingly, theories of grammar should not assume absolute homogeneity among speakers in order to be able to account for these idiolectal differences. On the other hand, the judgments of individual speakers on certain types of structures may vary from one particular instance of data collection to another. Linguistic theory should also be able to account for this kind of variation, and crucially, treat it as distinct from interspeaker variation.

The idea that the speaker-hearer's internal grammar is made up of a *core* and a *periphery*, also referred to as *grammatical indeterminacy*, is adopted by an increasing number of linguists (e.g., Bard, Robertson, & Sorace, 1996; Keller, 1998; Sorace, 1996; Sorace & Keller, 2005).[1] Pateman (1985) argues that the concept of "fuzzy" grammars and acceptability hierarchies is not necessarily incompatible with Universal Grammar (UG). He distinguishes between cases where UG: (a) strictly constrains the form of the grammar regardless of the input; (b) offers a preference for unmarked structures (over marked ones); and (c) does not prescribe an ordered set of structures for a small number of input patterns. Under this account, only (c) belongs to the periphery (Sorace, 1996). This position is appealing on empirical grounds, because it can explain why sentences such as (1b) and (1c) have an intermediate level of grammaticality. Yet, it has somewhat less predictive power than the traditional dichotomous models of grammaticality. That is, its ability to predict which sentences will have an intermediate level of grammaticality and on what basis sentences will vary along the grammaticality continuum is (at the present moment) questionable. Dichotomous models, on the other hand, attribute intermediate levels of grammaticality to the influence of extragrammatical factors on metalinguistic judgments (e.g., Bever, 1975). While these models make straightforward predictions with respect to the grammaticality status of sentences (i.e., grammatical or ungrammatical), researchers run the risk of falling into the traps of unfalsifiability

[1] It is not clear, however, whether the relationship between the core and the periphery is continuous or categorical (Sorace, 1996).

whenever they decide to attribute the marginal status of a sentence to extragrammatical factors, in part because it is impossible to eliminate these factors completely.[2] Until we have a better understanding of how they affect GJs, attributing intermediate levels of grammaticality to extraneous influences should be avoided (Schütze, 1996). For this reason, although no consensus has yet been reached on the nature of grammaticality, the former account is privileged.

### *What Do GJ Tasks Measure?*

***Grammatical competence vs. extragrammatical factors.*** The goal of contemporary linguistic theory is to describe grammatical competence, that is, the speaker-hearer's (subconscious) knowledge of the linguistic rules that constitute his or her internal grammar. This knowledge is to be distinguished from the speaker-hearer's actual use of the language, also referred to as performance (Chomsky, 1965). Grammatical competence is an abstraction and, as such, cannot be tapped directly; it can only be inferred from speaker-hearer's performance. One concern that arises in the literature is the extent to which GJs actually reflect grammatical competence. In the use of metalinguistic judgments, *grammaticality* and *acceptability* are often treated as synonyms, when they are in fact distinct concepts:

> The notion 'acceptable' is not to be confused with 'grammatical.' Acceptability is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence… Grammaticalness is only one of many factors that interact to determine acceptability. (Chomsky, 1965, p. 11)

Given this distinction, although the speaker-hearer has an intuitive[3] sense of grammaticality, his or her judgments can only be about the acceptability of linguistic structures.[4] Hence, the common misconception that GJ tasks provide a direct window into

[2] Note that researchers who consider grammaticality as gradient may still be caught up in unfalsifiable theories, insofar as grammatical indeterminacy does not necessarily entail the absence of extragrammatical factors. Nonetheless, this possibility is somewhat less likely, since a smaller portion of the intraspeaker variation would be attributed to such factors.

[3] A distinction is often made between *intuitions* and *judgments* (e.g., Schütze, 1996; Sorace, 1996). As per Chomsky's (1975) definition, intuitions pertain to the study of competence and, as such, are not directly observable, whereas judgments are (impure) performance data and can be observed.

[4] Note that for the sake of familiarity, and to ensure that acceptability would not be mistakenly interpreted as appropriateness, I will continue to use the term GJ.

linguistic competence (e.g., Carroll & Meisel, 1990; Ellis, 1990; Han, 2000) is clearly a fallacy: because grammaticality is not open to direct introspection, GJs are necessarily influenced by performance factors (e.g., Gass, 1994; White, 2003).

Sorace (1996), among others, identifies a number of extragrammatical factors that have been found to cause spurious judgments, namely: (a) parsing strategies, (b) context and mode of presentation, (c) pragmatic considerations, (d) linguistic training, and (e) language norms. Let us consider each of these factors. First, sentences involving parsing difficulties, such as the famous garden-path example *The horse raced by the barn fell*, tend to be judged as ungrammatical, even if they follow the wellformedness rules of the language (Bever, 1970). In this example, the past participle *raced*, which not only has the same phonetic form as its past-tense counterpart, but also is more frequently used intransitively, is more readily interpreted as the main verb of the sentence. Once the verb *fell* is encountered, the reader has to reanalyze the sentence to be able to parse the verb while following the wellformedness rules of the language. If the reader fails to recover from his or her initial analysis, the sentence is likely to be judged as ungrammatical. [5] Second, in addition to parsing strategies, the context in which sentences are presented can influence GJs. Levelt, van Gent, Haans, and Meijers (1977) suggest that GJs tend to be more positive for high imagery and concrete materials. Similarly, Greenbaum (1977), among others, shows that a sentence of marginal grammaticality is likely to be judged as grammatical if presented with sentences that are clearly ungrammatical, and as ungrammatical if presented with sentences that are clearly grammatical. Third, as Sorace (1996) explains, pragmatic factors can also have an effect on GJs. Specifically, Altmann and Steedman (1988) demonstrate that participants tend to go with the interpretation that is the most frequent and requires fewer assumptions about the previous discourse. Fourth, linguistic training have also been found to influence informants in their GJs. For example, Gleitman and Gleitman (1979) and Snow and Meijer (1977) show that informants with prior linguistic training differ in their GJs from informants without such training, although the nature of this difference is not clear. Finally, Sorace (1996) indicates that language norms also bias participants' GJs. For example, Greenbaum and

[5] Note that I use the word "reader," because garden-path effects often disappear in spoken language, as the listener can use prosodic information to eliminate temporary syntactic ambiguities (e.g., Kjelgaard & Speer, 1999; Schafer, Speer, & Warren, to appear).

Quirk (1970) suggest that GJs are influenced by the informants' beliefs about the forms that they normally use or that they should use, and by their willingness to tolerate forms that they neither use nor believe should be used. That is, informants might decide to rely on a different language norm than that underlying their internalized grammar. Hence, given the influence of extragrammatical factors on GJs, the idea that GJs provide a direct window into grammatical competence is very unlikely.

***The knowledge base underlying GJs.*** Another concern that arises in the use of GJ tasks is whether objectification skills such as metalinguistic judgments have anything to do with what is being objectified, here grammatical knowledge. That is, even if it was possible to eliminate the influence of extragrammatical factors on GJs, some researchers are convinced that GJs would still not reflect grammatical competence. Bever (1970), among others, argues that judgments (of any sort) are highly introspective and, as such, may be part of a general nonlinguistic cognitive system. The assumption that GJs are introspective is not atypical in the literature, as evinced by the use of think-aloud protocols in GJ tasks (e.g., Ellis, 1990, 1991). This procedure in general has shown a lack of internal consistency on the part of informants regarding the strategies they employ to determine grammaticality. These findings, however, should not be taken as evidence for the unsuitability of GJ tasks to investigate grammatical competence, but rather, as evidence for the inadequacy of introspective methods (such as the think aloud protocol) to examine grammatical competence.[6] Another argument against the use of introspection in general is that it is difficult, if not impossible, for people to be engaged in cognitive and metacognitive activities simultaneously (e.g., Vygotsky, 1979). For example, Ericsson and Simon (1984) have shown that in a text recall task, informants ceased to think aloud exactly when the difficulty level of the task increases. Hence, the above findings are expected if introspection (here, performed during the think aloud protocol) is indeed part of a general nonlinguistic cognitive system, as claimed by Bever (1970).[7] Yet, they do not entail that GJs themselves are introspective. Following Ringen (1977) and

[6] Introspective methods are justified if they are used to examine the informants' mental process, "independent of its veracity" (Schütze, 1996, p. 49).

[7] Basically, there is no reason to assume that informants should be homogenous if the strategies they are resorting to are not language-specific.

Pateman (1987), Schütze (1996) suggests that GJs are akin to physical sensations and, as such, cannot be verified:

> We can show that a person is lying about being in pain, but once we show that the report is sincere, no other evidence can disprove the assertion (it makes no sense to talk about the correctness of a sensation), and if two people's reports conflict in response to the same stimulus, *that does not undermine the veracity of either report* [emphasis added]. (pp. 50–51)

This analogy also entails that GJs, like physical sensations, may not always be accurate (Schütze, 1996).

There is no doubt that metalinguistic judgments differ from language use in a number of ways. For example, the former involves *controlled* processes, whereas the latter involves *automatic* processes (e.g., Ryan & Ledger, 1984). As a result, providing judgments is more difficult than using the language (i.e., in listening and speaking) (e.g., Gleitman & Gleitman, 1979). In addition, metalinguistic judgments exhibit more individual variation than linguistic capacity does (e.g., Hirsch-Pasek, Gleitman, & Gleitman, 1978). The fact that the ability to make metalinguistic judgments emerges later than the ability to use language is also indicative that the two processes are somewhat distinct (e.g., Hakes, 1980). Yet, Schütze (1996) argues that these differences do not necessarily entail that the knowledge base underlying GJs is *separate* from that underlying grammatical competence. Specifically, he considers the null hypothesis to be that GJs and grammatical competence are underlain by the same type of knowledge. This means that, until proven otherwise, GJs should be seriously considered as a potentially insightful source of information. It is, however, the task of researchers to find ways to minimize the effect of extragrammatical factors in order to extract from GJs only that information which pertains to grammatical competence.

## METHODOLOGICAL ISSUES

GJ tasks have been criticized not only because of the controversial status of metalinguistic judgments in linguistic theory, but also because they too often lack the rigorous control techniques normally used in psychological experimentation. This problem, which too often characterizes linguistic experimental methods, is summarized by Schütze (1996) as follows:

> …there are important shortcomings that arise because linguistic elicitation does *not* [emphasis in original] follow the procedures of psychological experimentation. Unlike natural scientists, linguists are not trained in methods for getting reliable data and determining which of two conflicting data reports is more reliable. In the vast majority of cases in linguistics, there is not the slightest attempt to impose any of the standard experimental control techniques, such as random sampling of subjects and stimulus materials or counterbalancing for order effects.
> (p. 4)

One of the main consequences of this lack of conscientious methodology is that little or no control is provided for the influence of extragrammatical factors on linguistic data. While it is obviously impossible to launder GJs from all impurities they may have, a number of basic measures can be taken to minimize the effect of performance factors on GJs.

### *Implementing Rigorous Experimental Designs*

***Materials.*** Schütze (1996) makes a series of methodological recommendations intended to minimize the influence of *test-related* and *subject-related* (extragrammatical) factors in GJ tasks. The first set of precautions concerns the design of materials used to elicit GJs. Since parsing difficulty has been found to influence GJs, researchers should avoid choosing sentences whose grammaticality rating could be confounded with parsing difficulty, unless the purpose of the experiment is to rate parsing difficulty. Because GJs may be also influenced by high imagery and concrete materials, sentences in context should not be compared to isolated sentences. Despite the lack of consensus among researchers, Schütze (1996) suggests that sentences be presented in context in order to reduce the likelihood that informants come up with their own imaginary context in which the sentence might occur. However, there are good reasons to question this recommendation, simply because the context itself can be an additional source of bias. This brings on the task of ensuring that informants judge a sentence as ungrammatical for structural reasons, and not because they believe it does not fit in the provided context.

Presenting sentences in isolation may thus provide a better control for the influence of pragmatic factors on GJs.

A third source of bias that Schütze (1996) discusses is the semantic content of the lexical items. Since GJs can be affected by imagery, lexical items should be varied to prevent this effect. Similarly, the frequency of lexical items should be controlled to avoid the possibility that informants reject sentences because of a word that are is not frequent in the language. Another factor which has been found to influence the informants' judgments is the order in which experimental items are presented (Greenbaum, 1973). Order effects can be caused by a number of factors, including nervousness at the beginning of the experiment, fatigue toward the end of the experiment, practice effects, and the influence of surrounding items. Fortunately, these effects can be neutralized by counterbalancing the order of experimental items across different partcipants. Schütze (1996) explains that the number of grammatical and ungrammatical items in the task can also influence GJs. For example, a greater number of grammatical items can lead informants to expect the test items to be grammatical and influence their judgments in general. Researchers should therefore ensure that the number of grammatical and ungrammatical items is more or less equal. Finally, in order to circumvent the possibility that subjects become aware of the purpose of the experiment, it is crucial to use *at least* as many distracters as experimental items.

***Procedures.*** The second set of recommendations that Schütze (1996) makes concerns the process of gathering GJs. First and foremost, GJ tasks should be administered to *random* samples of native speakers. This entails that informants should have no prior linguistic training. Unfortunately, this principle is the exception rather than the rule. Linguists have too often relied on their own judgments when creating theories of grammar. As Labov (1972) indicates, "linguists cannot continue to produce theory and data at the same time" (p. 199). As Schütze (1996) explains, informants should also be sufficient in number if any theoretical claim is to be made on the basis of GJ data. A large number of participants is necessary for the assumptions of statistical tests to be met and to ensure that the results are not distorted by the judgments of an atypical informant. Ideally, the sample should be as homogeneous as possible in terms of age, literacy, education, and idiolect, unless one wishes to examine the effect of these variables on

GJs.[8] Third, GJ tasks should be administered in a controlled setting, not only to reduce the chances that subject will become distracted during the task, but also to reduce between-subject variation by having all informants complete the task under the same conditions. In addition, Schütze (1996) emphasizes the importance of providing informants with instructions that are as precise and detailed as possible. Obviously, one cannot hope that the terms *grammatical* or *acceptable* would have the same meaning for different people. Instructions should provide the reasons why a sentence should be considered "good" or "bad," as well as the reasons that should not come into play in decision process. They should include examples that are unequivocally good or bad but not related to the constructions at issue. Instructions that are carefully formulated will help reduce the effect of pragmatic factors and language norms on the informants' judgments. That is, if informants are specifically told, for example, that a "good" sentence is a sentence that they could produce under certain circumstances (e.g., whether on the street or in a formal context), they will be less likely to judge a sentence that is *prescriptively* incorrect as "bad".

Another variable to consider in the design of GJ tasks which Schütze (1996) discusses in terms of whether the elicited ratings should be *absolute* or *comparative.* Absolute judgments can be binary (e.g., "sounds good" vs. "sounds bad") or involve a third, intermediate category (e.g., "not sure"). They can also include different levels of (un)grammaticality (e.g., "sounds okay" vs. "sounds good"; "sounds awkward" vs. "sounds bad"). By contrast, comparative judgments involve ranking sentences in comparison to one another on an acceptability scale (e.g., from "sounds good" to "sounds bad"). Schütze (1996) takes a neutral stand with respect to which type of rating GJ tasks should use. On the one hand, one can extract a lot of information from absolute judgments by converting them into rankings, whereas comparative judgments cannot be converted into absolute ratings (at least, conceptually). On the other hand, psychometric research shows that comparative ratings are more reliable than absolute ratings (e.g.,

[8] This seems to contradict the idea of random sampling. One problem in the use of GJ tasks with native speakers is that researchers tend to assume homogeneity and often limit their observations to group results. If individual results are not going to be considered, it is necessary to have a fairly homogeneous group in order to be able to make any theoretical claims, since GJs are likely to be influenced by age, literacy, education, and idiolect (see, for example, Birdsong, 1989, and Bialystock & Ryan, 1985, for review of studies examining the effect of literacy and education on GJs).

Mohan, 1977). In view of this trade-off, Schütze (1996) leaves it up to researchers to decide which type of rating is more appropriate to use. This will, in part, be determined by the position they adopt on the nature of grammaticality.

In contrast, Sorace (1996) favors the elicitation of comparative judgments. Her discussion is underlain by the assumption that grammaticality is a gradient concept. She argues that one of the problems with binary absolute judgments is that informants are forced to make a categorical decision on the grammaticality status of a sentence, when the sentence may in fact not be either completely grammatical or completely ungrammatical. One way to avoid this problem is to include several levels of (un)grammaticality in the rating scale. While including a middle, intermediate grammaticality level also appears to be a solution, the definition of this middle category is often problematic because it often confounds grammatical indeterminacy (i.e., the informant judges a sentence as being *between* grammatical and ungrammatical) and uncertainty (i.e., the informant *does not know* whether the sentence is grammatical or not). One way to circumvent this problem is to ensure that instructions include sufficient information about the meaning of that middle category so that it is not treated as uncertainty. It is also a good idea to include the category “don’t know” as a separate option in order to keep the two scales separate. In short, Sorace (1996) prefers ranking scales over rating scales because: (a) they are less constraining; (b) they have greater psychometric plausibility; and (c) they are more suitable to capture the notion of grammatical indeterminacy. From a theory-neutral perspective, however, it seems that either type of rating can be used, on condition that the shortcomings of absolute ratings as discussed above (e.g., limitations of binary scales) be taken care of.

After deciding which type of rating will be elicited, it is necessary to determine the amount of time informants will be given to provide their judgments. Schütze (1996) explains that limiting participants to only a certain amount of time has several advantages. First, initial reactions to given sentences are less likely to be influenced by extragrammatical factors such as pragmatic considerations and language norms. That is, if informants have less time to make a decision, they are less likely to think of sentences as appropriate or inappropriate and to consult their knowledge of prescriptive grammar. Second, time restriction makes it more difficult for informants to discover the structural analyses in which the researcher is interested. In other words, if subjects must provide an immediate response to sentences, they are less likely to figure out the purpose of the experiment. Another advantage of timed GJ tasks, which Schütze (1996) does not mention, is that since they are usually computer-administered, it makes it impossible for the informants to go back and modify earlier responses. Hence, in cases where *immediate* reactions are sought, controlling for the amount of time that subjects can take to make judgments is methodologically more suitable.[9]

***Analysis and interpretation of the results.*** Finally, the third set of precautions that Schütze (1996) provides concerns the analysis and interpretation of GJs. First, as in any psychological experiment, it is indispensable to perform statistical tests of significance on the results in order to establish whether the observed patterns (if any) should be attributed to grammatical competence (and error) or to error alone. If the tests are not significant, the researcher's task is to determine whether the data are faulty. While the above recommendations reduce the likelihood that GJs will be spurious, they do not in any way ensure that the obtained results will be free of impurities. Ultimately, GJs remain performance data from which grammatical competence can only be inferred. On the other hand, if the statistical tests are significant, the researcher should look for converging evidence before drawing generalizations from the results. One way to do this is through *cross-task validation*, that is, by comparing the results of the GJ task to those of other tasks and see whether they converge (e.g., Chaudron, 1983, 2003). If the results of the GJ task show up reliably across other types of tasks, researchers can be more confident that they represent fundamental underlying knowledge.

Furthermore, as Schütze (1996) explains, the results of GJ tasks should be replicable if the task is administered twice on the same sample (e.g., Bard et al., 1996; Chaudron, 1983, 2003; Gass, 1994; Schütze, 1996; Sorace, 1996). This criterion, also referred to as *test-retest reliability*, is especially important in the use of GJ tasks, because it has the potential to: (a) disentangle GJs that do not represent grammatical competence from those that do; and (b) shed some light on which linguistic structures belong of the core of grammar and which ones belong to the periphery, if one is to assume that grammaticality

[9] Note that timed GJ tasks are probably not appropriate for tasks testing the interpretive properties of given structures (e.g., scope quantification), since in such tasks, researchers are usually not interested in the informants' immediate reactions to sentences, but in their interpretation (Schütze, 1996).

is gradient (Sorace, 1996). With respect to (a), GJs on the same sentence type which vary considerably from one instance of data collection to another are probably not representative of grammatical knowledge, especially if the amount of time between the administration of the two tests is reasonably short (1–3 months).[10] As for (b), sentences that are given the same intermediate level of grammaticality on two separate instances of data collection might provide researchers with crucial information about the kind of linguistic structures that belong to the periphery of grammar. Replication is thus essential if one is to rely on GJs to make theoretical claims about grammatical competence. If *test-reliability* cannot be assessed due to time constraints or to the difficulty of having the same informants participate in the GJ task twice, *internal consistency* should at least be evaluated with the split-half reliability approach (in which test is divided into two parts and the correlation between the two halves is computed) or with reliability formulas such as the Kuder-Richardson 20 formula (K-R20) and the Kuder-Richardson coefficient 21 (K-R21) (for more details on these measures, see Brown & Hudson, 2002).

In light of the above discussion, the present study attempts to design and administer a GJ task that is not characterized by any of the theoretical and methodological problems associated with such tasks. This objective, albeit an ambitious one, is the first step toward the implementation of more rigorous experimental designs in the use of GJ tasks in linguistic theory. This study also aims to demonstrate that the FACETS software provides an efficient tool to analyze GJs and determine (a) which participants should be excluded from the analyses, (b) which test items should be revised, and (c) whether the results are reliable. Excluding participants who do not behave as expected (e.g., participants who did not do the task seriously) is necessary, as their GJs are unlikely to reflect their grammatical competence at all. Revising problematic test items is also important to increase the overall reliability of the task. Finally, assessing the reliability of the GJ task is essential in order to determine whether any firm conclusions can be drawn from the results. Given these objectives, the results of the GJ task will be analyzed from a methodological perspective only. That is, the implications of the actual GJs with respect

[10] One does not expect grammatical competence to vary much in so little time, at least with adult native speakers.

to the grammar of the speakers are not considered in this paper. For further detail on these implications, see Bullock, Omaki, Schulz, Schwartz, and Tremblay (2005).

## METHOD

### *Participants*

A total of 20 native speakers of American English (age 16–60, mean = 30) participated in the GJ task. They are undergraduate and graduate students at the University of Hawaiʻi at Mānoa who are majoring in different disciplines. Most of them (15/20) do not have any training in linguistics. The participants were paid $10 for their participation in this and two other tasks administered for a different research purpose (see Bullock et al., 2005 for details).

### *Procedures*

The GJ task used in this paper is part of a larger research project on the interaction between grammatical knowledge and processing constraints in Interlanguage (Bullock et al., 2005). The task was carried out in the Language Analysis and Experimentation Laboratories of the University of Hawaiʻi, which provide a controlled setting for the administration of language experiments. The task was administered on a computer with the software *E-Prime* (Psychology Software Tools, Inc., 2002). Sentences were presented one at a time on a computer screen, and for each them, the participants were given a maximum of 20 seconds to provide their judgment. The next sentence appeared on the screen immediately after the informants entered a response. If the participants did not make a judgment after 20 seconds, no response was recorded and the computer automatically moved on to the next sentence. This ensured that the participants would not take too much time to provide a response to the sentences. Given this mode of presentation, the informants were not able to go back and change their judgments on previous sentences.

The participants received very specific instructions as to the kind of judgments they were required to make for each sentence of the task (see Appendix I). The instructions first specified that the informants should not judge sentences on the basis of what they consider to be "proper English" or of what they were taught at school, but on the basis of whether they think they could or would say the sentence under appropriate circumstances. They were also told not to reject a sentence because they know of a better way to convey the same meaning. These specifications, for which concrete examples were provided, were made in an attempt to control for the possibility that sentences would be rejected on the basis of prescriptive rules of English, or because they were formulated in a way that was not necessarily the most natural for the informant. The instructions also specified the rating scale against which the participants were required to judge the sentences. The participants were told that a sentence should be judged as 'Perfect' (4) if they think the sentence sounds perfectly fine and they could or would say it under appropriate circumstances; a sentence should be judged as 'okay' (3) if they think the sentence is not completely perfect, but is still fairly good and that they might say it under appropriate circumstances; a sentence should be judged as 'awkward' (2) if the sentence sounds strange and they doubt they would ever say it; and a sentence should be judged as 'horrible' (1) if they think the sentences sounds terrible and they would never say it under any circumstance. Importantly, the participants were specifically told not to guess if they did not know or did not have any intuition regarding certain sentences, but instead, to select 'no intuition' (X). This enables the researchers to tease apart grammatical indeterminacy from uncertainty and increases the chance that the informants' judgments will indirectly reflect their grammatical competence.

### *Test Items*

The GJ task included a total of 80 items: 20 control items, 20 experimental items, and 40 distracters. The items were pseudo-randomized, and the order in which they were presented was counterbalanced across participants. The distracters used in this task serve as experimental items for a different study. For this reason, the judgments obtained on all the test items will be analyzed. The GJ task includes four general sentence types, which are broken down into four sentence subtypes with five tokens each (2)–(5):

(2) Tense-meaning (mis-)match, simple sentence (referred to as 'control'):

a. Isabelle is going to play tennis tomorrow.

b. Isabelle played tennis yesterday.

c. *Isabelle played tennis tomorrow.

d. *Isabelle is going to play tennis yesterday.

(3) Tense-meaning (mis-)match, complex sentence (referred to as 'experimental'):

a. John told Lucy yesterday that he is going to move to China.

b. John told Lucy that tomorrow he is going to move to China.

c. *John told Lucy tomorrow that he is going to move to China.

d. *John told Lucy that yesterday he is going to move to China.

(4) Embedded clauses with(out) subject-verb inversion, factive verbs (referred to as 'factive'):

a. John knew what Mary was doing.

b. Did John know what Mary was doing?

c. *John knew what was Mary doing.

d. *Did John know what was Mary doing?

(5) Embedded clauses with(out) subject-verb inversion, question verbs (referred to as 'question'):

a. Bill wondered where Keith is going shopping.

b. Did Bill wonder where Keith is going shopping?

c. *Bill wondered where is Keith going shopping.

d. *Did Bill wonder where is Keith going shopping?

These sentences are all fairly simple, because ultimately, the GJ task was designed to be conducted with second language learners of English. The control items are *simple* sentences in which the tense of the verb matches or does not match the meaning of the adverb. They are referred to as control items because in Bullock et al. (2005), they serve as a comparison point for the experimental items. The experimental items are *complex* sentences in which the tense of the verb matches or does not match the meaning of the adverb in either the main clause or the embedded clause. In particular, this sentence type tests whether informants know that the adverb in the main or embedded clause cannot modify the verb in the embedded or main clause, respectively. As their name suggests, the experimental items are the focus of the study carried out in Bullock et al. (2005). The factive items are declarative or interrogative sentences with a factive verb (e.g., *know*) in which the subject and the verb in the embedded clause have been inverted or not. Finally, the question items are declarative or interrogative sentences with a question verb (e.g., *wonder*) in which the subject and the verb in the embedded clause have been inverted or not. The factive and question items serve as distracters in Bullock et al. (2005).

### *Analyses*

Descriptive and comparative statistical analyses were conducted on the results using the program *SPSS* (2001). In addition, the ratings were analyzed with the software FACETS, a Rasch measurement computer program commonly employed in language testing research (Linacre, 1988). FACETS is generally used to analyze qualitative observations in the forms of ratings and ranks. By looking the interaction between several facets (i.e., variables) in the data, the program provides crucial information about the raters (here, the participants), the rated elements (here the sentences), and the internal consistency of the ratings. First, the program indicates whether the raters have a tendency to be lenient or harsh (here, to rate the test items as grammatical or ungrammatical, respectively) in their judgment. This kind of information is crucial in order to determine whether the participants have a general tendency to judge the test items as grammatical or ungrammatical, individually and as a group. The program also indicates whether sentence types tend to be rated leniently (here, as grammatical) or harshly (here, as ungrammatical). This information may help researchers revise the grammaticality status of certain sentences if they are consistently judged as more (or less) grammatical than what they expected. The program can also identify which individual test items are responsible for unexpected responses. This can help researchers determine whether any test items should be revised. In addition, the program assesses the internal consistency of the GJs by specifying, for example, which rater(s) and sentence type(s) exhibit more or less variation than expected (i.e., outliers and inliers). In the case of raters, this information is crucial for researchers to be able to exclude from the analyses the participants who have not completed the task very carefully. As for the sentence types, the researcher might want to consider revising individual test items within a given sentence type if the latter is found to exhibit more or less variation than expected (and if the individual test items within that sentence type are found to receive unexpected responses). For each of these analyses, FACETS provides a reliability estimate of the separation index, which determines whether the analysis can reliably distinguish between the different elements of a facet. Whether a high or low reliability estimate is desirable depends on the facet being examined. For each facet, the program also provides the *p* value of a fixed chi-square. The null hypothesis of the fixed chi-square is that all the elements of the facet are equal. A significant chi-square therefore leads to the rejection of the null hypothesis. Although some the analyses performed in FACETS can also be computed manually or with the help of other statistics programs, the FACETS program is a powerful tool because it performs all the calculations in a matter of seconds and provides researchers with both individual and group results that would otherwise be much more time-consuming to examine.

### *Adjustments to the Data*

Before proceeding to the results, it is necessary to mention the adjustments that have been made to the data. First, missing data (i.e., sentences for which no judgment was provided or for which the informants had no intuition) were replaced by the participant's mean on the grammatical or ungrammatical sentence type for which the data were missing. Second, an important typo was found in one of the tokens, which led half the participants to reject the sentence when it was meant to be grammatical. To maintain the number of tokens appearing in each condition equal, the score on this item was replaced by the participant's mean on that sentence type. Finally, ratings are examined on the 1–4 scale as opposed to being converted into dichotomous ratings (grammatical vs. ungrammatical), because (a) much information is lost in the process of dichotomizing, and (b) such dichotomies ignore the possibility that grammaticality is gradient, with sentences varying along the grammaticality continuum.

## RESULTS

### *Mean GJs*

Figure 1 summarizes the mean GJs for grammatical and ungrammatical control, experimental, factive, and question items. As shown in Figure 1, the effect of grammaticality was the greatest on control items, whereas ungrammatical factive and question items were not clearly judged as ungrammatical. A look at the individual ratings indicates that 14.8% (59/400) of the ungrammatical factive and question sentences were rated as 'Horrible' (1), 49.2% (197/400) as 'Awkward' (2), 18% (72/400) as 'Okay' (3), and 18% (72/400) as 'Perfect' (4).[11] The fact that 36% of the ungrammatical sentences were rated as grammatical explains why the mean is located close to the median (here, 2.5). The mean and standard deviation for each grammatical and ungrammatical control, experimental, factive, and question items are reported in Table 1. The standard deviations are all relatively low, with factive items showing the largest standard deviation. A 4 × 2 (Sentence Type × Grammaticality) repeated-measures ANOVA reveals a significant sentence type effect (*df* (3,19), *F* = 39.057, *p* < .001) and grammaticality effect (*df* (1,19), *F* = 531.600, *p* < .001), as well as a significant interaction between the two (*df* (3,19), *F* = 131.996, *p* < .001). This interaction can be observed in Figure 1 and Table 1, as the effect of grammaticality decreases from control sentences to experimental, factive, and question sentences.

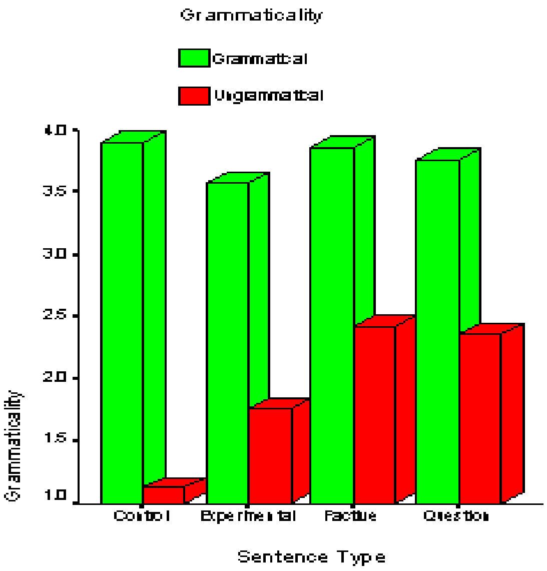

*Figure 1.* Mean GJs for grammatical and ungrammatical control, experimental, factive, and question items.

[11] The frequencies of ungrammatical factive and question sentences are combined because they are very similar.

Table 1

*Means and Standard Deviations for Grammatical and Ungrammatical Control, Experimental, Factive, and Question Items*

| Sentence Type | Mean | St. Dev. |
|---|---|---|
| Grammatical | | |
| Control | 3.90 | 0.22 |
| Experimental | 3.58 | 0.41 |
| Factive | 3.86 | 0.20 |
| Question | 3.76 | 0.27 |
| Ungrammatical | | |
| Control | 1.13 | 0.19 |
| Experimental | 1.76 | 0.40 |
| Factive | 2.43 | 0.61 |
| Question | 2.37 | 0.43 |

### *FACETS Analysis*

Figure 2 shows the multi-faceted ruler along which the participants, sentence type, grammaticality, and ratings are represented. In this output, the logit scale (–2 to 2 in this case) should be interpreted as representing an estimate of the true grammaticality status of sentences, with 0 being the cut-off point between grammatical and ungrammatical sentences. It can be observed from Figure 2 that the participants (101–120) form a normal distribution on the scale, with participant 103 being the most lenient rater (i.e., judging the test items more often as grammatical) and participant 118 being the harshest rater (i.e., judging the test items more often as ungrammatical).[12] Overall, the participants have a general tendency to judge the test items as grammatical, with 19 of them being represented above 0 on the logit scale. Factive and question items appear to be responsible for this trend, as they are located higher on the logit scale than experimental

[12] Recall that the logit scale as defined here represents grammaticality. This means that lenient raters are represented higher on the logit scale, and harsher raters are represented lower. This contrasts with the regular use of FACETS, in which the logit scale represents severity. On that scale, harsher raters are represented higher on the scale and lenient raters are represented lower.

and control items. This could mean that the researchers might need to label "ungrammatical" factive and question items as marginal (instead of ungrammatical), since they were not rejected as much as the researchers would have expected. Figure 3 shows the probability curves of each of the rating categories (1–4) in the task. Interestingly, these curves indicate that the participants did not use the intermediate grammatical rating ('Okay') as much as the intermediate ungrammatical rating ('Awkward'). In other words, when sentences were not considered strange in any way, they tended to be rated as 'Perfect'.

```
-----------------------------------------------
|Measr|+Participants       |-Type|-Gramm|S.1  |
-----------------------------------------------
+   2 +                    +     +      +(4)  +
|     |                    |     |      |     |
|     |                    |     |      |     |
|     |                    |     |      |     |
|     | 103                |     |      |     |
|     |                    |     |      |     |
|     |                    |     | Grammatical|
|     | 120                |     |      |     |
|     | 102  114           |     |      | --- |
|     | 116                |     |      |     |
+   1 +                    +     +      +     +
|     | 101  108  115      |     |      |     |
|     |                    |     |      |     |
|     | 105  110  111  117 |     |      |     |
|     | 119                | Factive    | 3   |
|     | 106  112  113      | Question   |     |
|     |                    |     |      |     |
|     | 104                |     |      |     |
|     | 107  109           |     |      |     |
|     |                    |     |      | --- |
*   0 *                    *     *      *     *
|     | 118                |     |      |     |
|     |                    |     |      |     |
|     |                    |     |      |     |
|     |                    | Experimental 2   |
|     |                    |     |      |     |
|     |                    |     |      |     |
|     |                    | Control    |     |
|     |                    |     |      |     |
|     |                    |     |      |     |
+  -1 +                    +     +      +     +
|     |                    |     |      |     |
|     |                    |     |      | --- |
|     |                    |     |      |     |
|     |                    |     | Ungrammatical
|     |                    |     |      |     |
|     |                    |     |      |     |
|     |                    |     |      |     |
|     |                    |     |      |     |
|     |                    |     |      |     |
+  -2 +                    +     +      +(1)  +
-----------------------------------------------
|Measr|+Participants       |-Type|-Gramm|S.1  |
-----------------------------------------------
```

*Figure 2.* Multi-faceted ruler along which the participants, sentence types, grammaticality, and ratings are represented.

```
 -2.0             -1.0              0.0              1.0              2.0
  ++----------------+----------------+----------------+----------------++
1 |                                                                     |
  |                                                                     |
  |                                                                     |
  |                                                                  444|
  |                                                              4444   |
P |11111                                                      444       |
r |     111                                                444          |
o |        111                                          444             |
b |           111                                     44                |
a |              111                                44                  |
b |                 111                           44                    |
i |                    11   22222               44                      |
l |                 22222***      222222      44                        |
i |            22222        111        2222444                          |
t |       22222                11        44222                          |
y |  22222                       111   4*33333****33333                 |
  |22                            333***3          222  33333333         |
  |                         33333444   1111          2222      33333333 |
  |                  3333333 4444          1111          222222        3|
  |      33333333333*44444444                  111111111       222222222|
0 |******44444444444                                    1111111111111111|
 ++----------------+----------------+----------------+----------------++
 -2.0             -1.0              0.0              1.0              2.0
```

*Figure 3.* Probability curves of the ratings.

Table 2 reports the FACETS analysis for each of the participants who completed the GJ task, including the grammaticality logit, the standard error of the logit, the mean-square infit statistic, and the mean-square outfit statistic. The infit statistic is the weighted mean-squared residual which is sensitive to irregular inlying patterns, whereas the outfit statistic is the unweighted mean-squared residual which is sensitive to unexpected outliers (Linacre, 1989–1996). In the context of a GJ task, informants whose infit or outfit is lower than 0.5 exhibit too little variation in their GJs, whereas informants whose infit or outfit is higher than 1.5 show excess variation in their GJs.[13] The results are reported in order of decreasing grammaticality.

[13] I follow Lunz and Stahl (1990) in using 0.5 and 1.5 as cut-off points to identify misfitting participants.

Table 2

*Fit Statistics for Each Participant*

| Participant | Grammaticality (logits) | Error | Infit | Outfit |
|---|---|---|---|---|
| 103 | 1.56 | .16 | 0.8 | 1.2 |
| 120 | 1.33 | .16 | 1.0 | 1.1 |
| **102** | 1.23 | .16 | 0.9 | **1.6** |
| 114 | 1.18 | .16 | 0.8 | 0.6 |
| 116 | 1.05 | .16 | 1.1 | 1.1 |
| 108 | .92 | .16 | 0.6 | 0.5 |
| 115 | .90 | .16 | 0.8 | 0.9 |
| 101 | .87 | .16 | 0.8 | 1.2 |
| 117 | .72 | .16 | 0.8 | 0.9 |
| 105 | .72 | .16 | 0.8 | 0.9 |
| 111 | .69 | .16 | 0.5 | 0.5 |
| 110 | .69 | .16 | 0.5 | 0.5 |
| **119** | .64 | .16 | **2.0** | **2.7** |
| 112 | .53 | .16 | 0.8 | 0.9 |
| **113** | .51 | .16 | 1.2 | **1.9** |
| 106 | .48 | .16 | 0.9 | 1.3 |
| 104 | .27 | .16 | 0.6 | 0.5 |
| **109** | .21 | .16 | 1.5 | **2.3** |
| **107** | .21 | .16 | **2.1** | **2.4** |
| 118 | –.05 | .16 | 0.9 | 0.8 |
| Mean | .73 | .16 | 1.0 | 1.2 |
| Std. Dev. | .40 | .00 | 0.4 | 0.6 |

Notes: Reliability of separation index = .84; fixed (all same) chi-square: 123.4, *df* = 19, significance: *p* < .00

As shown in Table 2, the standard error of the logit is fairly low for all the participants (.16). The infit and outfit statistics indicate that there are five misfitting informants in the data (participants 102, 107, 109, 113, 119). That is, the GJs of five participants are found not to be internally consistent. Of these, two participants (107 and 119) have *both* an infit and an outfit higher than 2.0, which means that their responses are not be predicted by the model. Interestingly, these two participants were also found at the time of the data collection to have completed the task noticeably faster than the others, which suggests that they might not have answered the GJ task very carefully. In view of the fact that FACETS analyzed these two informants as misfitting, and that the researchers had initial doubts on whether to include their GJs in the analyses, the data from participants 107 and 119 should be excluded from the analyses. As for the remaining misfitting participants, in a larger sample, one should consider excluding them from the analyses, but given the small size of this sample, it might not be necessary to do so.[14]

Table 2 also reports the reliability estimate of the separation index, which indicates the extent to which the analysis distinguishes between the informants with respect to their tendency to judge sentences as grammatical or ungrammatical. In this case, a low reliability estimate is desirable, in the sense that participants should ideally exhibit the same or similar judgments on the same sentence types. As the analysis shows, the reliability estimate for the participants is in fact very high ($r$ = .84), which means that the informants were not equally lenient in their judgments on the same sentence types. The high $p$ value of the fixed chi-square ($p$ < .00) confirm that these results are statistically significant. In other words, it indicates that the null hypothesis (i.e., that all participants are equal in their GJs) must be rejected.

Table 3 reports the FACETS analysis for each sentence type in the GJ task, including the grammaticality logit, the standard error of the logit, the mean-square infit statistic, and the mean-square outfit statistic. The results are reported in order of decreasing

[14] FACETS works best with large samples of raters. The smaller the sample, the more misfitting participants will be found. For this reason, caution is advised in excluding participants from small samples solely on the basis of the fit statistics, especially if only one of the fit statistics (i.e., infit or outfit) is below 0.5 or over 1.5. In the present case, participants 107 and 109 were excluded, because *both* the infit and outfit statistics were extremely high, but also, because the FACETS analysis coincided with the initial doubts of the researchers concerning participants 107 and 119.

grammaticality. As can be seen in Table 3, the standard error of the logit for all sentence types is quite low (.07-.08). The infit and outfit statistics indicate that none of the four sentence types tested in this study are misfitting. This means that the GJs obtained on each sentence type are predictable and internally consistent. Table 3 also reports the reliability estimate of the separation index, which indicates the degree to which the analysis distinguishes between the four sentence types. In this case, a high reliability estimate is perhaps desirable, as the sentence types tested are not necessarily expected to be judged grammatically equally. This is in fact what is found (*r* = .98). The high *p* value of the fixed chi-square (*p* < .00) confirms that these results are statistically significant. This means that the null hypothesis (i.e., that all sentence types are rated equally) must be rejected.

Table 3

*Fit Statistics for Each Sentence Type*

| Sentence Type | Grammaticality (logits) | Error | Infit | Outfit |
|---|---|---|---|---|
| Question | .62 | .07 | 1.0 | 1.4 |
| Factive | .47 | .07 | 0.9 | 1.4 |
| Experimental | –.37 | .08 | 1.3 | 1.3 |
| Control | –.72 | .07 | 0.7 | 0.6 |
| Mean | .00 | .07 | 1.0 | 1.2 |
| Std. Dev. | .56 | .00 | 0.2 | 0.4 |

Notes: Reliability of the separation index = .98; Fixed (all same) chi-square: 242.3, *df* = 3, significance: *p* < .00

Table 4 reports the FACETS analysis for grammaticality (i.e., grammatical vs. ungrammatical items), including the grammaticality logit, the standard error of the logit, the mean-square infit statistic, and the mean-square outfit statistic. Again, the results are reported in order of decreasing grammaticality. As shown in Table 4, the standard error of the logit is for grammatical and ungrammatical sentences is again quite low (.07 and .04, respectively). The infit and outfit statistics reveal that although raters judge grammatical sentences with more variation than ungrammatical sentences, the pattern of responses is predictable. The GJs for grammatical versus ungrammatical sentences can therefore be considered internally consistent. Table 4 also reports the reliability estimate of the separation index, which specifies the extent to which the analysis distinguishes between grammatical and ungrammatical sentences. Here, a high reliability estimate is clearly desirable, as grammatical sentences should be judged differently from ungrammatical ones. This is exactly what is found ($r = 1.00$), and the high $p$ value of the fixed chi-square ($p < .001$) confirms that these results are statistically significant. In other words, the null hypothesis (i.e., that grammatical and ungrammatical sentences are rated equally) must be rejected.

Table 4

*Fit Statistics for Grammaticality*

| Sentence Type | Grammaticality (logits) | Error | Infit | Outfit |
|---|---|---|---|---|
| Grammatical | 1.41 | .07 | 1.2 | 1.5 |
| Ungrammatical | –1.41 | .04 | 0.9 | 0.9 |
| Mean | .00 | .06 | 1.0 | 1.2 |
| Std. Dev. | 1.41 | .01 | 0.1 | 0.3 |

Notes: Reliability of the separation index = 1.00; Fixed (all same) chi-square: 1209.8, *df* = 1, significance: $p < .00$

Finally, although none of the sentence types were found to be misfitting, it is always a good idea to examine individual test items to determine whether any of them behaves differently from the remaining test items within the same sentence type. Table 5 reports each of the individual test items that received an unexpected GJ by three or more participants, including the sentence type to which it belongs, its expected score, and its mean residual. The mean residual refers to the mean difference between the observed score and the expected score on the test item. As shown in Table 5, only four test items received an unexpected GJ by three or more participants. Noticeably, these items are all grammatical factive and question sentences that were judged significantly less grammatical than expected. Of these items, three were interrogative sentences and three had the verb *wonder*. Importantly, the two test items that received the greatest number of unexpected responses are the declarative sentence *Bill wondered where Keith is going shopping* and its interrogative counterpart *Did Bill wonder where Keith is going shopping?* At this point, while it is not clear why the remaining two items caused spurious judgments, the number of participants who provided an unexpected GJ to these two sentences (6 and 8, respectively) is high enough for the researchers to revise these sentences before future administrations of the test. The researchers might also want to consider eliminating these items from the results, as they were found to behave significantly differently from the remaining test items within the same sentence type.

Table 5

*Individual Test Items that Received Three (or More) Unexpected Responses*

| | Sentence Type | Unexpected Responses | Expected Score | Mean Residual |
|---|---|---|---|---|
| Did Mary guess where the party was? | Factive | 3 | 3.9 | –1.9 |
| Bill wonder where Keith is going shopping. | Question | 6 | 3.9 | –1.2 |
| Did Bill wonder where Keith is going shopping? | Question | 8 | 3.9 | –1.5 |
| Did Lucy wonder what the book was about? | Question | 3 | 3.9 | –1.5 |

## DISCUSSION

This paper first aimed to design and administer a GJ task that is not characterized by the problems commonly associated with such tasks. This objective was only partially met. On the one hand, the test items used in the GJs included 20 control sentences and 20 experimental sentences out of a total of 80 sentences. The ratio between the number of experimental items and the remaining test items (1:3) was high enough for the informants not to be able to know what the researchers were trying to assess. The experimental items were pseudo-randomized and the order in which they were presented was counterbalanced across the participants, which ensured that no order effect would be found in the results. The instructions were very specific and teased apart grammatical indeterminacy from uncertainty. They provided concrete examples of the kind of information that the researchers are interested in (e.g., errors violating the morphosyntactic rules of English) and the kind of information that should not come into play in the decision process (e.g., language use not following the prescriptive rules of English). All the participants completed the GJ task in the same controlled setting, which was essential to minimize the influence of extraneous factors. The task was computer-administered and left the participants with a maximum of 20 seconds to provide a judgment on each sentence. This ensured that the participants would not take too long to make a decision, and it also made it impossible for them to go back and change their answers. Hence, the GJ task itself can be considered methodologically-sound, as it follows a number of measures intended to minimize the influence of extragrammatical factors on the GJs. On the other hand, the number of participants recruited for the task was small. As a result, it is not clear that the sample met the assumptions underlying the statistical analyses used in this paper. Since FACETS works best with a large sample of informants, the results reported in this study perhaps displayed more variation than necessary.[15] Another limitation of this study concerns the broad age range of the participants, which suggests that the sample tested might not be as homogeneous as one would like it to be. Given these two shortcomings, any conclusions drawn on the basis of this study with respect to the grammar of the participants should be made with caution.

This study also aimed to demonstrate that the software FACETS provides an efficient tool to analyze GJs and determine (a) which participants should be excluded from the analyses, (b) which test items should be revised, and (c) whether the results are reliable. This objective was met. Using FACETS as a method of analysis, it was found that the GJs of one fourth of the participants (5/20) may not be internally consistent. The infit and outfit statistics were particularly high for two of these participants. Independent evidence gathered during the data collection process suggested that the two participants in question

[15] Since the sample appears to be normally distributed, a larger sample might still show a number of misfitting raters, but they are likely to represent a smaller percentage of the total sample than in the present study (Martyn K. Clark, personal communication, June 6, 2005).

did not do the GJ task carefully. In light of these results, it was highly recommended that these two participants be excluded from the analyses. As for the remaining misfitting participants, given the relatively small size of the sample, it was decided that it was not necessary to eliminate their results. Using FACETS, it was also found that the participants tended to judge sentences as more grammatical than the researchers would have expected. The sentence types that appeared to be responsible for this asymmetry are the factive and question sentence types. On the basis of these results, it was suggested that the researchers consider "ungrammatical" factive and question items as marginal (instead of ungrammatical). Interestingly, the test items which received the greatest number of unexpected responses also belonged to the factive and question sentence types. In view of this finding, it was proposed that two of the four test items that generated 6 and 8 unexpected responses be revised before future administration of the test, and that they perhaps be eliminated from the results since they were found to behave significantly differently from the remaining test items within the same sentence type. Finally, despite these problematic test items, the FACETS analyses conducted on the sentence types and on grammaticality showed that the GJs provided on each sentence type are internally consistent. These results suggested that the researcher could rely on the GJs to support previously-formulated theoretical claims, while bearing in mind the limitations of the sample size.

## CONCLUSION

The use of GJ tasks is essential in linguistic theory because it can provide crucial information about grammatical competence that elicited production tasks and naturalistic data collection cannot offer. If designed and administered carefully, GJ tasks can provide empirical evidence that may serve in the formulation, support, and refinement of theoretical claims in the study of language. The present paper provided a general overview of the questions surrounding the use of GJ tasks in linguistic theory and suggested solutions to overcome their methodological flaws. It stressed the importance of clearly defined theoretical assumptions and rigorous experimental designs in the use of GJ tasks as a data-collection method. Following these recommendations, a GJ task was designed and administered to a group of native speakers of American English. FACETS was introduced as a means for analyzing the ratings, and the analyses generated by the program were used to determine (a) which participants should be excluded from the analyses, (b) which test items should be revised, and (c) whether the GJs should be considered reliable. Using FACETS as a novel method for the analysis of GJs proved essential to solving some of the methodological problems commonly associated with GJ tasks, namely the lack of appropriate statistical measures assessing the reliability of GJs. On the basis of the fit statistics provided in this paper, it was determined that the results were internally consistent and that conclusions could reliably be drawn from the study, while keeping in mind the relatively small size of the sample. This study, albeit not a perfect one, represents a first step toward the implementation of methodologically sound GJ tasks.

Future research should be conducted in three directions. First, it should aim to determine precisely how extragrammatical factors influence GJs. Understanding the effect of these factors appears to be the only possible way to extract from GJs only that information which pertains to grammatical competence. Second, it should follow Schütze’s (1996) recommendations regarding the creation of test items, the specification of procedures, and the analysis and interpretation of the results in the design and administration of GJ tasks. Finally, the use of FACETS should become a standard in the analysis of GJs, as it provides a very efficient tool for scrutinizing the data and determining whether any firm conclusions can be drawn from the results.

## ACKNOWLEDGMENTS

This research was partially supported by the Elizabeth Holmes-Carr scholarship awarded to Gerald Bullock, Akira Omaki, Barbara Schulz, Bonnie D. Schwartz, and the author of this paper. Many thanks to James D. Brown, Martyn K. Clark, and Tomomi Hasegawa for their valuable comments on this study.

## REFERENCES

Altmann, G. T. M., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition, 30*, 191–238.

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

Bever, T. G. (1970). The cognitive base for linguistic structures. In R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Bever, T. G. (1975). Functional explanations require independently motivated functional theories. In R. E. Grossman, J. San, & T. J. Vance (Eds.), *Papers from the parasession on functionalism* (pp. 580–609). Chicago: Chicago Linguistic Society.

Bialystock, E., & Ryan, E. B. (1985). A metacognitive framework for the development of first and second language skills. In D.-L. Forrest-Presley, G. E. MacKinnon, & T. G. Waller (Eds.), *Metacognition, cognition, and human performance. Volume 1: Theoretical Perspectives* (pp. 207–252). Orlando: Academic.

Birdsong, D. (1989). *Metalinguistic performance and interlinguistic competence*. New York: Springer-Verlag.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing.* Cambridge, MA: Cambridge University.

Bullock, G., Omaki, A., Schulz, B., Schwartz, B. D., & Tremblay, A. (2005). *Did they say in Interlanguage that CP is missing, or did they say that in Interlanguage CP is missing?* Paper presented at the 8th Bi-Annual Generative Approaches to Language Acquisition Conference, Università degli Studi di Siena, September 9.

Carroll, S., & Meisel, J. (1990). Universals and second language acquisition: Some comments on the state of current theory. *Studies in Second Language Acquisition, 12*, 201–208.

Chaudron, C. (1983). Research on metalinguistic judgments: A review of theory, methods, and results. *Language Learning, 33*(3), 343–377.

Chaudron, C. (2003). Data collection in SLA research. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 762–828). Oxford: Blackwell.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT.

Chomsky, N. (1975). *The logical structure of linguistic theory*. New York: Plenum.

Chomsky, N. (1991). Universal grammar. Letter to the editor. *The New York Review of Books, 38*(21), 82.

Chomsky, N. (1995). *The Minimalist Program*. Cambridge, MA: MIT.

Ellis, R. (1990). Grammaticality judgments and learner variability. In H. Burmeister & P. Rounds (Eds.), *Proceedings of the 10th Second Language Research Forum* (Vol. 1, pp. 25–60). Eugene, OR: University of Oregon.

Ellis, R. (1991). Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition, 13*(2), 161–186.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports on data*. Cambridge, MA: MIT.

Gass, S. M. (1994). The reliability of second-language grammaticality judgments. In E. E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second language acquisition* (pp. 303–322). Hillsdale, NJ: Lawrence Erlbaum Associates.

Gleitman, H., & Gleitman, L. R. (1979). Language use and language judgment. In C. J. Fillmore, W. Kempler, & W. S.-Y. Wang (Eds.), *Individual differences in language ability and language behavior* (pp. 103–126). New York: Academic.

Goss, N., Zhang, Y.-H., & Lantolf, J. P. (1994). Two heads may be better than one: Mental activity in second-language grammaticality judgments. In E. E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second language acquisition.* (pp. 263–286). Hillsdale, NJ: Lawrence Erlbaum Associates.

Greenbaum, S., & Quirk, R. (1970). *Elicitation experiments in English: Linguistic studies in use and attitude*. Coral Bables, FL: University of Miami.

Greenbaum, S. (1973). Informant elicitation of data on syntactic variation. *Lingua, 31*, 201–212.

Greenbaum, S. (Ed.). (1977). *Acceptability in language*. The Hague: Mouton.

Haegeman, L., & Guéron, J. (1999). *English grammar: A generative perspective*. Malden, MA: Blackwell.

Hakes, D. T. (1980). *The development of metalinguistic abilities in children*. New York: Springer-Verlag.

Han, Y. (2000). Grammaticality judgment tests: How reliable are they? *Applied Language Learning, 11*(1), 177–204.

Hirsch-Pasek, K., Gleitman, L. R., & Gleitman, H. (1978). What did the brain say to the mind? A study of the detection and report of ambiguity by young children. In A. Sinclair, R. J. Jarvella, & W. J. M. Levelt (Eds.), *The child's conception of language* (pp. 97–132). Berlin: Springer-Verlag.

Keller, F. (1998). Gradient grammaticality as an effect of selective constraint re-ranking. In M. C. Gruber, D. Higgins, K. S. Olson, & T. Wysocki (Eds.), *Papers from the 34th meeting of the Chicago Linguistic Society. Vol 2: The panels* (pp. 95–109). Chicago: Chicago Linguistic Society.

Kjelgaard, M. M., & Speer, S. R. (1999). Prosodic facilitation and interference in the resolution of temporary syntactic closure ambiguity. *Journal of Memory and Language, 40*, 153–194.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania.

Levelt, W. J. M., van Gent, J. A. W. M., Haans, A. F. J., & Meijers, A. J. A. (1977). Grammaticality, paraphrase, and imagery. In S. Greenbaum (Ed.), *Acceptability in language* (pp. 87–101). The Hague: Mouton.

Linacre, J. M. (1988). Facets: A computer program for many-facet Rasch measurement [Computer software]. Chicago: Mesa.

Linacre, J. M. (1989–1996). *A user's guide to Facets: Rasch measurement Computer Program*. Chicago: Mesa.

Lunz, M. E., & Stahl, J. A. (1990). Judge severity and consistency across grading periods. *Evaluation and the Health Professions, 13*, 425–444.

Mohan, B. A. (1977). Acceptability testing and fuzzy grammar. In S. Greenbaum (Ed.), *Acceptability in language* (pp. 133–148). The Hague: Mouton.

Pateman, T. (1985). From nativism to sociolinguistics: Integrating a theory of language growth with a theory of speech practices. *Journal for the Theory of Social Behaviour, 15*, 38–59.

Pateman, T. (1987). *Language in mind and language in society: Studies in linguistic reproduction*. Oxford: Clarendon.

Psychology Software Tools, Inc. (2002). E-Prime (Version 1.1) [Computer software]. Pittsburgh: Author.

Ringen, J. D. (1977). On evaluating data concerning linguistic intuition. In F. R. Eckman (Ed.), *Current themes in linguistics: Bilingualism, experimental linguistics, and language typologies* (pp. 145–160). Washington: Hemisphere.

Ryan, E. B., & Ledger, G., W. (1984). Learning to attend to sentence structure: Links between metalinguistic development and reading. In J. Downing & R. Valtin (Eds.), *Language awareness and learning to read* (pp. 149–171). New York: Springer-Verlag.

Schafer, A. J., Speer, S. R., & Warren, P. (to appear). Prosodic influences on the production and comprehension of syntactic ambiguity in a game-based conversation task. In M. K. Tanenhaus & J. C. Trueswell (Eds.), *World situated language use: Psycholinguistic, linguistics and computational perspectives on bridging the product and action tradition.* Cambridge, MA: MIT.

Schütze, C. T. (1996). *The empirical base of linguistics: Grammaticality judgments and linguistics methodology*. Chicago: The University of Chicago.

Snow, C. E., & Meijer, G. (1977). On the secondary nature of syntactic intuitions. In S. Greenbaum (Ed.), *Acceptability in language* (pp. 163–177). The Hague: Mouton.

Sorace, A. (1996). The use of acceptability judgments in second language acquisition research. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 375–409). San Diego, CA: Academic.

Sorace, A., & Keller, F. (2005). Gradience in linguistic data. *Lingua, 115*(11), 1497–1524.

SPSS Inc. (2001). SPSS: Real Stats, Real Easy (Version 11.0) [Computer software]. Chicago: Author.

Vygotsky, L. S. (1979). Consciousness as a problem of psychology of behavior. *Soviet Psychology, 17*, 3–35.

White, L. (2003). *Second language acquisition and Universal Grammar*. New York: Cambridge University.

## APPENDIX I

### Instructions

In this task, you will be asked to judge whether individual sentences sound good or not TO YOU. A sentence sounds good if you think you would or could say it under appropriate circumstances. By contrast, a sentence sounds bad if you think you would never say it under any circumstances. Here, we are interested in your linguistic intuitions only, not in the rules of "proper English" that you might have been taught in school. For example, consider sentence (1):

(1) Mary never goes nowhere.

In the past, you might have been taught that sentences with double negatives such as (1) are not acceptable in "proper English". However, under certain circumstances (e.g., if you are angry), you may actually produce these sentences yourself. If this is the case, then you should NOT judge (1) as bad. Basically, we would like you to ignore the language rules that you might have been taught in the past and focus strictly on whether you think you would or could say the sentence under appropriate circumstances. Now, compare sentence (1) to sentence (2):

(2) *Peter drinked the entire bottle of wine.

In contrast to sentence (1), as a native speaker of English, you would never say sentence (2), because 'drinked' sounds bad in English under ANY circumstance. This means that you should judge (2) as bad.

Sometimes there is more than one way to express meaning. For example, consider sentences (3) and (4):

(3) John gave Mary the book he bought at the bookstore after his class yesterday.

(4) John gave the book he bought at the bookstore after his class yesterday to Mary.

Here, you might prefer (3) over (4), simply because 'to Mary' in (4) is so far from 'gave' that sentence (4) does not sound as natural as (3). This, however, does not imply that sentence (3) is necessarily bad. When judging sentences, you should focus on whether a sentence sounds good to you, not on whether there is a better way to convey the same meaning with a different sentence.

Last but not least, because we are interested in your linguistic intuitions, it is ESSENTIAL that you do NOT spend a lot of time trying to figure out what linguistic rules might be violated in the sentence. We are interested in your IMMEDIATE reaction to the sentences. Taking too much time to provide a judgment will distort our findings and invalidate our results.

*Procedures*

You are about to read 80 individual sentences. For each sentence, you need to indicate whether the sentence sounds perfect, okay, awkward, or horrible to you. In order to indicate your response, press the number on the keyboard that corresponds to each judgment:

| perfect | okay | awkward | horrible |
|---|---|---|---|
| 4 | 3 | 2 | 1 |

If you think a sentence sounds perfectly fine and would or could say it under appropriate circumstances, select the option 'perfect' (4). If you think the sentence is not completely perfect but it is still pretty good and you might say it under appropriate circumstances, select the option 'okay' (3). On the other hand, if you think the sentence sounds strange and you doubt you would ever say it, select 'awkward' (2). Finally, if you think the sentence sounds terrible and you would never say it under any circumstance, select 'horrible' (1).

If you do not know or do not have any intuition, DO NOT GUESS. Instead, indicate that you do not have any intuition by pressing X on the keyboard:

| No intuition |
|---|
| X |

It is CRUCIAL that you do not guess when you have no intuition, because this will distort our findings and invalidate our results.

You have a maximum of 20 seconds to make your judgment. If you make your judgment before the end of the 20 seconds, you will automatically move on to the next sentence. If you have not made a judgment by the end of the 20 seconds, you will also move on to the next sentence. It is thus imperative that you do not take too much time to make your judgment.

Do you have questions before you start the task? If you have doubts about how to complete the experiment, please tell the researcher right now. Otherwise, you are now ready to begin. Have fun!