A Solution toward Bettering the Validity of the Language Acceptability Assessment Elicitation Procedure

Ludovic M. Kovalik
Winston-Salem State University

1. The language acceptability assessment elicitation: Benefits and drawbacks

In days of old, the decisions concerning the acceptability of language structures tended to mirror the workings of the minds of linguists—more often than not, linguists speculating about language while dwelling in ivory towers, far off the typical venues of real language: the mouth and the brain of the everyday speaker. The advent and rise of psycholinguistics heavily challenged that pattern, and most linguists have come to accept the fact that valid judgments relative to the acceptability of a language structure can only be obtained "out there," in the midst of ordinary speakers (Deese, 1970; Garnham, 1985; Hatch, 1983; Kess, 1976; Schütze, 1996). Consequently, the language acceptability assessment elicitation procedure started being seen as a useful instrument in linguists' attempt to reveal how language works.

As do most novel things, the language acceptability assessment elicitation has alternated periods of extreme popularity with periods of near-oblivion. Several accounts have been given for such a state of affairs. Insofar as the periods of sustained usage of the instrument are concerned, they are linked mainly to the fact that usage thereof may be quite helpful in researchers' efforts to collect information about structures that are not frequently encountered in language. Such information is most reliable when the structure whose acceptability needs to be assessed is presented in context. That is particularly so because judges assessing a sentence or a structure in isolation attempt to call to mind contexts wherein that sentence or structure could occur (Gass, 2001; McDaniel and Cairns, 1996; Nagata and Bain, 2000; Schütze, 1996; van Dijk, 1977). Given that some judges might be more successful than others in their endeavor to call up a context possibly validating said sentence/structure, a lot of unwanted variation might arise in the assessments. In contrast, in-context presentation of the sentence or structure to be assessed is likely to lead to more inter-rater consistency, thus increasing the reliability of the judgments.

As for the periods of lesser interest in the language acceptability assessment elicitation procedure, they seem to be mostly caused by two shortcomings thereof—the "artificiality" of the situation and the low generalizability of its findings, respectively. As emphasized by Schütze (1996), the first of those backlashes refers to the fact that the procedure requires that judges display a sort of behavior different from their everyday behaviors. However, Schütze indicates that some people consider making judgments of others' discourse as playing a fairly central role in our everyday lives, given that we ceaselessly assess others' utterances in order to establish, for instance, whether or not we belong with the same community. What Schütze fails to emphasize is that our judgments of others' discourse are likely to be done unawares and non-prescriptively in everyday life situations—which is definitely not the case in elicitation tasks, where assessments tend to be conscious and prescriptive.

The second of the backlashes of the procedure mentioned in the literature has to do with the low level of generalizability of its results. This is particularly so because, in their attempt to provide "good" answers, people might give high acceptability ratings to structures they dislike (or do not use) under normal circumstances while being harsh in their assessment of structures highly accepted in their discourse community; given that, researchers relying entirely on the

elicitation procedure might end up drawing false generalizations concerning the discourse perceptions of their raters (Maynor, 1982). Nevertheless, as indicated by Schütze (1996), there is a relatively simple solution to that problem, in that eliciting judgments from a large number of respondents will render vain individual judges' attempts to cosmeticize their discourse usage or perception, thus leading to enhanced generalizability of findings.

While the literature mostly focuses on the two pitfalls of the acceptability assessment elicitation procedure mentioned above, there exist other dangers associated with the instrument, as well. The present paper addresses one such potential drawback—namely, the possibility that the judges' ratings reflect not the degree of acceptability of the structure assumedly being assessed but some other concerns the raters might have, rather, such as punctuation, word order, the likeability of neighboring structures, and sometimes even the very likeability of the message conveyed by the text in which the structure to assess occurs.

If that is so, the validity of the findings arrived at via acceptability assessment elicitation is heavily challenged. Fortunately, nevertheless, there is a solution to that problem—a solution that is detailed upon in the remainder of this paper.

2. An acceptability assessment elicitation

The potential lack of validity of findings obtained through language acceptability assessment elicitation and a solution for overcoming that problem were prompted by the experiment described below.

2.1. The experiment: Questionnaire, participants, and requirements

In broad terms, the experiment underlying this paper was about assessing the acceptability of a number of sentences. Some of those sentences were full-fledged, others were elliptical, with ellipsis possibly affecting either a NP, or a VP, or a NP-VP complex, and still others were grammatically ill-constructed (the latter category of sentences is irrelevant to this paper, so no further reference will be made to it). The experiment involved more than mere assessment of the acceptability of certain structures, nevertheless. Indeed, in addition to rating the degree of acceptability of a number of sentences presented to them on a survey form, the participants (N=172, all native speakers of American English) had a second task to fulfill—namely, they had to provide a betterment suggestion for each of the sentences they deemed as less than altogether acceptable; the specifics concerning how those suggestions were to be made are presented below.

All the sentences to be assessed were presented in context, the decision for in-context presentation thereof being accounted for by the researcher's creed that the most reliable acceptability ratings of structures are made where context is provided. In order to make it clear to the judges that they were to rate the acceptability of *a certain* sentence within each of the stretches of text on the questionnaire, the how-to-complete-the-survey instructions explained to them that their assessments were to reflect the degree of acceptability of the sentence in bold in each of those texts. Such clarification was deemed as necessary because, as emphasized in the literature (see, for instance, Douglas, 2000), it is of extreme importance that judges have a clear picture of what they are expected to do. Another thing pointed out in the instructions was that the acceptability ratings were to be made using the 4-point scale provided underneath each of the items, with the specification that a rating of 4 was to be given to a sentence deemed as totally acceptable and ratings of 3, 2, and 1—to sentences deemed as somewhat acceptable, somewhat

unacceptable, and totally unacceptable, respectively. A further thing that the instructions taught the judges was how to make betterment suggestions. Specifically, it was explained to them that if they considered that a word or group of words should be omitted from a sentence, they were to strike through that word or group of words; in contrast, if they were of the opinion that a word or group of words should be added to a sentence, they were to insert that word or group of words in the space between the lines.

As mentioned above, the questionnaire involved a number of sentences—some elliptical in nature, others full-fledged—that had to be rated for acceptability. Firstly, it involved 12 elliptical sentences, all of which had been identified in one of two genres of written discourse: the short story and the breaking news article. The short story genre was represented by John Steinbeck's "The Chrysanthemums" and Ray Bradbury's "The Crowd," and the breaking news genre was represented by a number of articles (taken from *The New York Post*, *The New York Times*, or *The Washington Post*) tackling one of two quite like events: the 2000 presidential elections in the USA and Romania, respectively. Secondly, the questionnaire carried researcher-generated ellipsis-free rewritings of those elliptical sentences, the reason why ellipsis-free rewritings of originally elliptical sentences were favored over non-elliptical sentences lying in the researcher's hope that preserving all the discourse features of a sentence other than ellipsis would diminish, if not altogether eliminate, the danger of extraneous variables coming into play in the assessment making—e.g, text intelligibility or word order.

The elliptical sentences to be rated for acceptability carried several subtypes of ellipsis. One subcategory of sentences involved nominal ellipsis[1] only, with one or several NP(s) left out of the discourse—see example (1) below. The second subcategory of sentences involved verbal ellipsis only, with one or several verb(s) undergoing ellipsis—see example (2) below. The third subcategory of sentences on the questionnaire involved NP-VP ellipsis [2] only, with one or several complex(es) made up of a NP and a VP affected by ellipsis—see example (3) below. Finally, the fourth subcategory of sentences to be rated involved two discourse elements fulfilling different syntactic functions affected by ellipsis within one and the same sentence (henceforth, "multiple ellipsis"); an instance of such a sentence, involving a subject ellipsis and a subject-and-verb ellipsis, is presented in (4) below.

(1) Elisa's lips moved silently, forming the words "Good-bye—good-bye." Then she whispered, "That's a bright direction. There's a glowing here." The sound of her whisper startled her. She took herself free and looked about to see whether anyone had been listening. Only the dogs had heard. **They lifted their heads toward her from their sleeping in the dust, and then stretched out their chins and settled asleep again**.

(2) Elisa stood in front of her wire fence watching the slow progress of the caravan. **Her shoulders were straight, her head thrown back, her eyes half-closed, so that the scene came vaguely into them**.

---

[1] Ellipsis can affect a NP fulfilling *any* syntactic function. However, only ellipses of nominals functioning as subject were identified in the corpus from which the samples of real discourse presented to the raters were taken; as a result, the terms "nominal ellipsis" and "subject ellipsis" as used henceforth should be seen as interchangeable.

[2] All the NPs in the complex ellipses submitted for assessment were subjects, rather than objects; hence the occasional label "subject-and-verb ellipsis" assigned to such ellipses hereafter.

(3) Poverty and crime were the main issues for voters Sunday in Romania, where the favorites in the presidential election, an ex-Communist who held the job for six years and an outspoken nationalist, both vowed to stamp out corruption. President Emil Constantinescu did not seek a second term. Dissatisfaction with his leadership has set the stage for a comeback by Ion Iliescu, a 70-year-old former Communist who was president from 1990 to 1996. **Iliescu has accused the five-party ruling coalition of ruining the economy, vilified privatization and said the country should not rush to enact reforms needed to gain entrance into the European Union and NATO**.

(4) The man climbed over the single-tree, steadying himself with a hand on the burro's white rump. He settled himself in the seat, picked up the lines. "Thank you kindly, ma'am," he said. "I'll do like you told me; I'll go back and catch the Salinas road." He clucked his tongue. The beasts leaned luxuriously into their collars. **The wagon turned and crawled out the entrance road and back the way it had come, along the river**.

As mentioned above, the questionnaire did not involve elliptical sentences, only, but ellipsis-free rewritings thereof, as well. The bolded sentence in (5) below illustrates a non-elliptical rewriting of an originally verb-elliptical sentence. Note that, for emphasis' sake, the originally ellipted verb form *are* in the ellipsis-free rewriting is given in italics here; needless to say, it was not italicized on the questionnaire.

(5) Florida, with its 25 electoral votes, is the only state that matters now, as the nominee who carries it will win the election. **This campaign will spawn dozens of what-might-have-been scenarios, but few *are* more poignant than this**: Had Gore carried his home state of Tennessee, he'd be the president-elect, with or without Florida.

There were two criteria underlying the selection of the discourse samples to be included in the questionnaire. One of those criteria stemmed out of the desire to ensure variety across genres and sentence types. As for the second criterion, it had to do with the need to ensure that enough discourse was provided for the judges to understand the texts and correctly interpret the ellipses therein, if any were present. As a result, the questionnaire items were quite varied in terms of length, ranging from two lines of text (19 words) in the case of a discourse sample coming from Bradbury to over half a page of text (201 words) in the case of one coming from a breaking news article on the presidential elections in Romania—a topic the judges were not expected to know much about.

2.2. Research questions

There were four research questions that the experiment was hoping to provide responses to, but only two of those are relevant here. The first of those questions was meant to indicate whether there was a difference between the degree of acceptability, as reflected in the judges' ratings, of an elliptical sentence presented in context and the ellipsis-free rewriting of that sentence. As for the second of the two questions relevant to this study, it was meant to indicate whether, in the case of an elliptical sentence deemed as not altogether acceptable, it was the elliptical nature of said sentence that accounted for its less-than-total acceptability or some other reasons(s), rather.

2.3. Results and discussion

In order to obtain a response to the first of the two research questions relevant here (and thus to establish whether there were statistically significant relationships between the acceptability ratings of the elliptical sentences and those of the non-elliptical ones), the ratings given by the judges to the bolded sentences on the questionnaire were entered in the computer and interpreted using SPSS. The findings arrived at are summed up in Table 1 below.

Sentence type	Mean rating	S.D.	t-test
Elliptical	3.04	.458	$t=5.432$
Non-elliptical	3.16	.400	$p < .05$

Table 1: Paired-samples t-test by sentence type

As shown by Table 1, the paired samples t-test brought to the fore a statistically significant relationship between the acceptability ratings obtained for the elliptical sentences and those obtained for the non-elliptical rewritings of those sentences ($t=5.432$, $p<.05$), the elliptical sentences being rated as significantly less acceptable than the non-elliptical ones. Based on the findings displayed in the table, the conclusion apparently lending itself was that the participants in the study (and, possibly, native speakers of American English, in general) tended to favor full-fledged structures over elliptical ones.

However, as indicated earlier in this paper, the elicitation focused upon four distinct subcategories of elliptical structures and their non-elliptical rewritings. Given that the elliptical sentences and their ellipsis-free counterparts did not differ greatly in terms of mean ratings and standard deviations (witness the means of 3.04 and 3.16 and the standard deviations of .458 and .400, respectively), it was deemed worthwhile seeing whether the elliptical sentences were rated consistently lower than their non-elliptical rewritings. As a result, a second set of statistics was run, the findings of which are presented in Table 2 below.

Sentence		Mean rating	S.D.	t-test
Focus	Type			
Nominal	Elliptical	3.00	.831	$t=1.376$ n.s.
	Non-elliptical	3.09	.804	
Verbal	Elliptical	3.34	.502	$t=1.106$ n.s
	Non-elliptical	3.30	.529	
Complex	Elliptical	3.00	.552	$t=6.478$ $p < .05$
	Non-elliptical	3.23	.414	
Multiple	Elliptical	2.90	.721	$t=1.322$ n.s.
	Non-elliptical	2.82	.685	

Table 2: Paired-samples t-test by sentence type and focus

As seen in Table 2, the sentences involving ellipsis of nominal elements, verbal elements, and multiple elements, respectively, had acceptability ratings highly comparable to those of their non-elliptical counterparts—in fact, the ratings for the verb-elliptical sentences and the multiple-elliptical sentences were even slightly higher than those for the non-elliptical rewritings. Such a finding strongly suggests that the judges in this study (and, possibly, native speakers of American English, in general) perceive nominal-, verbal-, and multiple-elliptical sentences and full-fledged ones as equally acceptable. In contrast, the acceptability ratings for the complex-elliptical sentences were significantly lower than those for their non-elliptical counterparts (t=6.478, p<.05); that appears to suggest a marked dislike, on the part of the study participants (and, maybe, on the part of native speakers of American English, as such) of ellipses of subject-and-verb complexes.

The findings displayed in the two tables above appeared to answer the first of the research questions mentioned earlier—i.e, the question relative to whether there was a difference between the degree of acceptability, as reflected in the judges' ratings, of an elliptical sentence presented in context and the acceptability of the ellipsis-free rewriting of that sentence. Based on the findings hitherto presented, the response to that question appeared to be that while most of the elliptical structures under analysis were as accepted by the judges as were non-elliptical ones, complex-elliptical sentences were deemed as less acceptable than non-elliptical ones. However, whether or not that response was valid was to be shown by a minute analysis of the betterment suggestions made by the 172 raters for the sentences deemed as less-than-altogether acceptable.

Said analysis involved two steps, each involving two substeps. The first of the two steps translated as establishing the total number of ratings given for the complex-elliptical sentences and their non-elliptical rewritings, on the one hand, and as determining the number of ratings other than 4 'totally acceptable' for those sentences, on the other. As for the second of the two steps, it entailed identifying the judges who had given said sentences ratings lower than 4, on the one hand, and reading/classifying their betterment solutions, on the other. The results of the analysis are presented in Table 3 below.

Sentence type	Total ratings	Ratings of 4	Betterment solutions			
			expected	missing	given	addressing the S-V complex
Elliptical	852	329	523	92	431	127 (29%)
Non-elliptical	852	369	483	72	411	111 (27%)

Table 3: Complex-elliptical sentences vs. non-elliptical rewritings

As indicated by Table 3, 852 ratings were given for each of the two sentence types discussed here—complex-elliptical vs. non-elliptical. Of the 852 ratings given for the elliptical structures, 329 were ratings of 4 'totally acceptable' and 523 were ratings lower than 4. As a result, there should have been 523 betterment solutions for the elliptical sentences; however, only 431 such solutions were offered, of which 127 (i.e., 29%) translated as inserting the elliptical subject-and-verb complex. The numerical findings were pretty much the same for the ellipsis-free rewritings of the originally complex-elliptical sentences: a total of 852 ratings, of which 369 ratings of 4

and 483 ratings lower than 4; 411 betterment suggestions made, of which 111 (i.e., 27%) translated as deletion of the researcher-added subject-and-verb complex.

The most interesting of the findings presented in Table 3 is that the percentages of betterment suggestions addressing the subject-and-verb complex were highly comparable—note that 29% of the solutions given for the complex-elliptical sentences advocated insertion of a left-out subject-and-verb complex, which is quite similar to the 27% of the solutions given for the non-elliptical rewritings advocating deletion of an overtly-present subject-and-verb complex. Such a finding appears to suggest that just as they saw no difference between the degrees of acceptability of subject-, verb-, and multiple-elliptical sentences, on the one hand, and full-fledged sentences, on the other, the participants in the study saw no difference between subject-and-verb-elliptical sentences and full-fledged sentences, either.

If so, why was it that the complex-elliptical sentences had significantly lower mean acceptability ratings than their non-elliptical counterparts, nevertheless? A bulletproof answer to that question would be hard, if at all possible, to provide, given the intuitive nature of assessment-making, on the one hand, and the impossibility for one to "read" the judges' minds after the fact, on the other. However, it became evident during the analysis of the betterment suggestions offered that many of the ratings had nothing to do with the nature (elliptical vs. non-elliptical) of the sentences being assessed—in fact, many of the ratings other than 4 'totally acceptable' had to do with extraneous matters, such as word order, words/phrases in the text not related to ellipsis (or overt presence) of a subject-and-verb complex, difficulty of the text, and even the rater's familiarity with (or acceptance of) the message conveyed in the text. Indeed, quite a few bettered rewritings of the less-than-altogether-acceptable sentences identified during the analysis suggested that a pronoun in the text be replaced with a full noun (or the other way around), and others suggested replacement of an infinitive in the text with a gerund (or vice versa). A further category of acceptability-enhancing suggestions advocated addition of an adverb (or even a conjunction) to, or removal thereof from, the text, and yet another one raised the issue of sentence length, recommending splitting up a sentence deemed as too long into two. Moreover, a few of the bettered rewritings even revolved around the importance of *Mr.*, which is often omitted in newspaper style, in front of a male's last name.

The betterment suggestions above, all of which were found in two or more respondents' survey forms, were not the only ones offered by the questionnaire takers. In fact, in addition to the above solutions, there were quite a few other ones (found in one answer sheet, only); those solutions are not mentioned here, given that providing a full list of suggestions made by the judges with a view to bettering the acceptability of the sentences on the questionnaire is deemed irrelevant. What matters, rather, is that even though the complex-elliptical sentences had a significantly lower mean acceptability rating than their full-fledged rewritings, that statistically significant difference does not appear to have had to do with the elliptical nature of said sentences—recall that only 29% of the suggestions relative to the complex-elliptical sentences translated as inserting the elliptical subject-and-verb complex, which is quite similar to the 27% of the solutions advocating removal of the overtly present subject-and-overt complex from the ellipsis-free rewritings of the elliptical originals.

It follows then, based on the analysis of the betterment suggestions, that the ratings made by the participants in the study might not have addressed what they were expected to address—namely, the degree of acceptability of the elliptical sentences as compared with that of the full-fledged ones. Rather, most of the ratings appear to have revolved around issues other than the nature (elliptical vs. non-elliptical) of the sentences whose acceptability had to be rated, thus

rendering the validity of the findings of the elicitation highly questionable. Indeed, had it not been for the requirement that the judges offer betterment solutions for the sentences they deemed as less-than-altogether acceptable, (some of) the elliptical structures under investigation could have appeared to be less accepted by them (and, by extension, by native speakers of American English, in general) than they actually are. That might have led to faulty conclusions regarding the acceptability of ellipsis in written English discourse, thus jeopardizing the validity of the language acceptability elicitation procedure as such.

Based on the study described here, it would be safe to conclude, therefore, that the validity of the language acceptability assessment elicitation procedure is not beyond doubt. However, said validity can be increased by incorporating into the questionnaire at the core of the procedure the request that judges offer a betterment solution for each of the structures deemed as less-than-altogether acceptable.

3. Conclusions

The two main characteristics of the language acceptability assessment procedure described above were that the sentences that required rating for acceptability were presented in context and that the judges were asked to provide betterment solutions for all the sentences perceived as less-than-altogether acceptable. In-context presentation of the sentences to be rated was deemed as important in that the degree of acceptability of a structure is context-dependent, rather than set once and for all. As for the request for betterment suggestions, it was accounted for by the researcher's interest in finding out what lay behind a rating of 'less than acceptable.' Indeed, whereas one would expect a bettered rewriting of an ungrammatical sentence to address the ungrammaticality therein, an acceptability-enhancing rewriting of a grammatically well-formed sentence, be it elliptical or full-fledged, is a lot harder, if at all possible, to foresee.

The experiment indicated that grammatically-valid structures, whether elliptical or non-elliptical in nature, tended to be deemed by the judges as equally acceptable. Indeed, three of the four subtypes of elliptical sentences—namely, the subject-elliptical, the verb-elliptical, and the multiple-elliptical ones—were rated as equally acceptable as their non-elliptical rewritings. However, a statistically significant difference was found between the mean rating of the complex-elliptical sentences (i.e, sentences involving ellipsis of a subject-and-verb complex) and that of their full-fledged counterparts. Such a difference, which appeared to challenge the truth-value of the conclusion relative to the similarity in the ways elliptical sentences and full-fledged sentences are perceived, called for a careful analysis of the betterment suggestions made for the complex-elliptical sentences and their non-elliptical rewritings. That investigation revealed that when a sentence is deemed as not altogether acceptable, its less-than-total acceptability is not normally accounted for by its nature (elliptical vs. non-elliptical) but, rather, by issues having nothing to do with said nature—e.g, punctuation, word order, content of the text, difficulty of the text, and even likeability of the text.

Such a conclusion was only possible due to the fact that the experiment leading to it involved not only assessing the acceptability of certain structures but also offering betterment solutions for the structures deemed as less-than-altogether acceptable. Indeed, but for the specific request that the raters offer a solution for each of the sentences perceived as not quite acceptable, the validity of the findings of the study might have been doubtful. That should be seen as strongly advocating the need to associate language acceptability assessment elicitations with the

requirement that betterment solutions be offered by the raters for any text/structure that is given a rating other than 'totally acceptable.'

References

Deese, James. (1970). *Psycholinguistics*. Allyn and Bacon, Boston.

Douglas, Dan. (2000). *Assessing languages for specific purposes*. Cambridge University Press, Cambridge, MA and New York.

Garnham, Alan. (1985). *Psycholinguistics: Central Topics*. Methuen, London and New York.

Gass, Susan M. (2001). Innovations in second language research methods. *Annual Review of Applied Linguistics* 21: 221-232.

Hatch, Evelyn M. (1983). *Psycholinguistics: A second language perspective*. Newbury House, Rowley, MA.

Kess, Joseph F. (1976). *Psycholinguistics: Introductory perspectives*. Academic Press, New York.

Maynor, Natalie. (1982). Grammatical judgments in *LANE*. *American Speech* 57: 228-231.

McDaniel, Dana and Helen S. Cairns (1996). Eliciting judgments of grammaticality and reference. In Dana McDaniel, Cecile McKee, and Helen S. Cairns (eds.), *Methods for assessing children's syntax*. MIT Press, Cambridge, MA. 233-254

Nagata, Hiroshi and Bruce Bain. (2000). Judgments of grammaticality of sentences with a differing number of arguments: A comparison of English and Japanese speakers. *Perceptual and Motor Skills* 91: 503-511.

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

van Dijk, Teun A. (1977). Acceptability in context. In Sidney Greenbaum (ed.), *Acceptability in language*. Mouton, The Hague, Paris, and New York. 39-61.

Author contact information:

Ludovic M. Kovalik: kovaliklm@wssu.edu