Journal of Education and Human Development
December 2016, Vol. 5, No. 4, pp. 222-231
ISSN: 2334-296X (Print), 2334-2978 (Online)

Published by American Research Institute for Policy Development
DOI: 10.15640/jehd.v5n4a21
URL: https://doi.org/10.15640/jehd.v5n4a21

The Effect of Response Format on Test Performance: Problems and Possible Solution (An Alternative Grammaticality Judgement Rating Scale)[1]

Naif Alsaedi[2]

Abstract

Intuitional data elicited by means of grammaticality judgement (GJ) tasks are affected by a diverse number of linguistic and non-linguistic factors including the type of the measurement scales and the response formats used (see Schütze, 1996; Sorace, 1996). The problems associated with the rating scales vary according to the type of the rating scale used - whether absolute or comparative. Such problems pose a serious challenge to linguists to find an alternative rating scale that can overcome these methodological problems and pitfalls. The only way to do so is by using a new rating scale which allows sharp lines to be drawn between the learner's certainty, doubt and lack of knowledge reflected in his or her judgements. The four-point scale in its new format proposed in this article (i.e., *clearly correct*, *clearly incorrect, possibly incorrect* and *do not know*) managed to map the territory between the three possibilities that capture a learner's feelings towards any given sentence. What is unique about this rating scale compared to others commonly used in grammaticality judgments (i.e., Coppetiers, 1987, Schachter and Yip, 1990; Schachter, 1990; Gass, 1994) is how it works and how the data obtained by means of this scale can be marked and scored. The purpose of the present article is to introduce this scale and to discuss from a logical point of view the extent to which it can produce reliable and valid data that reflect second language (L2) learner's interlanguage knowledge.

Key Words: grammaticality judgement, acceptability, rating scale, reliability, second language, interlanguage, Universal Grammar

1. Grammaticality Judgement Tests in Second Language Acquisition Research

Native mastery of a language such as English refers to the learner's subconscious acquisition of a complex internal system of finite syntactic structures and rules. The system that forms the English grammatical competence (the knowledge of syntactic structures and rules) of the native speaker, also referred to within the universal grammar (UG) research as the "I-language" (Chomsky, 1986, p. 23), allows the speaker to produce an infinite number of possible utterances and to intuitively reject the impossible ones. Therefore, for researchers to obtain information about the defining characteristics of the syntactic structures that form the grammar of a particular I-language, they must make inferences about what native speakers believe to be grammatical/acceptable and ungrammatical/unacceptable in the known language.[3]

[1] I wish to express my deepest gratitude to Professor Martha Young-Scholten and Professor Anders Holmberg for many fruitful discussions and remarks; however, no one but me can be held responsible for remaining errors. Also, I gratefully acknowledge the financial support I have received from Taibah University.
[2] e-mail: n.s.alsaedi@newcastle.ac.uk
[3] The notion of acceptability differs from the notion of grammaticality; acceptability refers to the native speaker's or L2 learner's intuitional judgement about a sentence, whereas grammaticality is a theoretical term used by linguists or grammarians to establish whether a sentence conforms to the requirements of the grammar of the given language. Chomsky (1965) also stated that "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" (p. 11). For more detailed information about the different between these two notions, refer to Haegeman (1994) and Adger (2003).

All native speakers have a built-in ability to automatically and correctly judge which sentences are and are not well formed in their native language. To successfully construct native-like grammar, an L2 learner must acquire similar ability – competence – enabling him or her to determine the well-formedness of sentences in the target language. Therefore, GJ tests have been widely used since the mid-1970s basically by UG-based L2 researchers such as White (1985, 1986, 1992), Liceras (1989), Schachter (1989), and Pérez-Leroux and Glass (1997). These and other researchers have widely used GJ tasks to test their theoretical claims and to examine the grammatical competence of L2 learners. This is because such elicitation tests make it possible for researchers "to investigate aspects of inter-language [L2 linguistic system] which may not otherwise be amenable to inspection" (White, 2003, p. 18). Tremblay (2005, p. 159) argues that GJ tasks can provide crucial information about grammatical competence that elicited production tasks and naturalistic data collection cannot offer''. These intuitional tasks accomplish this by placing certain constraints on the learner so that he or she is "forced to make choices within a severely restricted area of his [or her] phonological, lexical or syntactic competence" (Corder, 1973, p. 41, cited in Gass, 1994, p. 306). In this way, the L2 learner cannot avoid considering the specific structure under investigation as he or she may do freely in the unconstrained production data elicitation tasks. This point to another advantage of using GJ tasks in generative linguistics–related research. Because the learner is required only to judge the given sentences and not generate or articulate them, some performance factors such as slips of the tongue, unfinished utterances, speech misinterpretations, and others that may hide some properties of the underlying competence are eliminated.

However, the use of GJ tests in second language acquisition (SLA) research has been a subject of debate since the eighties; the serious issue at the centre of this ongoing debate among linguists concerns the validity of grammaticality judgments – that is, whether the data obtained from these kinds of tests are reliable or not in reflecting L2 learners' linguistic competence (e.g., Bley-Vroman, Felix, and Loup, 1988; Birdsong, 1989; Cook, 1990; Carroll and Meisel, 1990; Ellis, 1991; Gass, 1994; Davies and Kaplan, 1998; Tremblay, 2005). In other words, these and other researchers have criticised the use of GJ tests on the grounds that gathered GJ data are necessarily affected by a diverse number of task-related and subject-related factors that could invalidate or at least bias the obtained results unless they are properly controlled for.[4] For example, one of the important task-related factors in grammaticality that has been a matter of concern for some researchers is the measurement scale - the rating scales - used for judgment elicitation. The following section aims to investigate the effect of response format of GJ tests on L2 learners'; in other words, it will discuss the influence response format might leave on the reliability of the results of GJ tests

2. Rating scales: their effect on L2 learners' judgment behaviour

Since ''grammaticality judgments are to be considered empirical data in the sense of experimental psychology, then the measurement scale used for judgment elicitation is of crucial importance'' (Keller, 1999, p. 000). Birdsong (1989) argues that researchers in order to investigate the reliability of GJ tasks need to investigate the impact of response format on L2 learners' performance on these tests. Ellis (1991) for example observes that the performances of L2 learners are often inconsistent on GJ tasks; he argues that the stability of their responses changes depending on the response type. In what follows, I will illustrate meticulously how the properties of the current response format used in the current GJ tests affect the reliability of the data obtained from these kinds of tests. However, for the sake of illustration, it is useful to base the critical discussion here on a few studies that use grammaticality judgements as an elicitation task.

[4] Discussing all of these factors is an issue beyond the scope of this paper, which is intended to address only the effect of measurement scales on judgment behaviour. However, the interested reader is referred to Schütze (1996), where she discusses the major test-related and subject-related factors (such as the learner's level of proficiency, the amount of time given for the task, the mental state of the individual making the judgement, knowledge about specific language rules, and the learner's test-taking strategies, such as guessing) and proposes a set of methodological recommendations that could increase the reliability of the collected data.

A binary, 2-point scale (e.g., *grammatical* vs. *ungrammatical*) is usually avoided in SLA research.[5] This is because L2 learners are usually asked to make judgements during stages in which their knowledge about certain elements of the target grammar is incomplete or even totally absent; this status is typically referred to in the literature as "indeterminacy".

Indeterminacy in a learner's developing grammar has led some linguists (e.g., Coppetiers, 1987; Bley-Vroman, Felix, and Loup, 1988) to include a third intermediate option in their scales: usually, the *not sure* selection. The analysis of this third option can be problematic, however. While Coppetiers's solution is to consider *not sure* responses as meaning "correct", Bley-Vroman et al. consider these responses as meaning "incorrect". Both solutions create serious methodological problems that can affect the reliability of the scores. One of these problems concerns equalizing certainty with uncertainty (or doubt) regarding the student's feelings about the judged sentences. In other words, judging a sentence as *grammatical* or *ungrammatical* reflects a high degree of confidence about it, while judging it with *not sure* reflects much less confidence. Keeping in mind that the goal of psycholinguistic research is to measure interlanguage (IL), this forces us to ask what these *not sure* responses really reveal to us about grammar development. Other groups of researchers who have used 3-point scales have presented another solution to this analysis problem by excluding all of the *not sure* responses from the analysis. This solution, however, even if it may produce more reliable scores than the other two do, still falls short of producing accurate data, raising the same question as with regard to Coppetiers's and Bley-Vroman's solutions. What, exactly, do *not sure* responses tell us about grammar development? Furthermore, another serious problem can arise when respondents give too many *not sure* responses.

Such problems and pitfalls associated with the 3-point scale have led some linguists (e.g., Schachter and Yip, 1990; Schachter, 1990) to use a 4-point scale ranging from, for example, *clearly correct* to *possibly incorrect*, in agreement with researchers who treat acceptability as a gradient concept (cf. Sorace, 1996; Tremblay, 2005) in the sense that sentence acceptability depends *partly* on the strength of the learner's preference regarding how to say it.[6] Yet the results analyses gathered using these scales are still problematic. Schachter (1990), for example, counted the *possibly correct* option as if it were the *clearly correct* option and the *possibly incorrect* as if it were the *clearly incorrect* option, allowing for no distinction between the *possibly* and *clearly* categories. Conversely, Schachter and Yip's (1990) results reflect the separateness of these four options. Such unjustified analyses force us to consider the value of including options that cannot tell us anything about the learner's IL since the options are not part of the analysis in any real way, such as in Schachter's study. How do we, then, interpret the two possibly (in) correct options in light of Schachter and Yip's analysis? In addition, what if none of the options applies, such as in the case where the given sentence is beyond the level of the learner, and therefore not part of his or her IL?

The same problems arise with five-point and other multipoint scales used in GJ tasks. In addition, these scales place a greater difficulty on the learner to choose from multiple options in a task that is already highly effortful. Consider the following 7-point scale used in Gass (1994), where she asked her participants first to judge categorically the (un)grammaticality of test sentences and then assess their degree of confidence or doubt regarding each sentence they judged:[7]

–3	–2	–1	0	+1	+2	+3
Definitely incorrect			unsure			definitely correct

It is very possible that some (if not all) L2 learners find it difficult to differentiate between the middle options of Gass's scale – for example, between +2 and +1 or –2 and –1, or even between –1, 0 and +1.

[5] Such two-response scales are relatively common in first language syntactic research.

[6] More importantly, acceptability of a sentence depends on other factors, including the sentence's grammaticality, the context in which it is uttered, and whether it is difficult to parse. Thus, it can be said that not every sentence, even if it is well formed, is considered acceptable by all learners (or even by all native speakers of that language), and not every ungrammatical sentence is considered unacceptable by all learners.

[7] The scope of this section does not allow for further discussion of the problems connected with all of the multipoint scales used in SLA research.

Compared to other scales, however, data obtained from such multipoint scales would produce more complications. Such complications emerge from the same questions raised above with Schachter and Yip's (1990) and Schachter's (1990) studies, but in greater degree: How should a linguist interpret each symbol (number) that indicates more or less acceptability compared to the other ones?

More precisely, how is it possible to map territories between all the symbols, especially the ones in the middle? Do these obtained scores reflect the learners' abstract grammatical knowledge? And what can such a score tell us in general about the process of SLA?

I find it difficult to interpret a learner's judgements on sentences from a multipoint confidence scale in a way meaningful enough to increase our knowledge about the process of SLA.[8] In addition, I think it is extremely difficult for an L2 learner to judge sentences on a multipoint scale based on his or her initial feelings, even when asked to do so. I can imagine that the participants in Gass's study discussed above reflected for a moment on which option best described their feelings towards the sentences before they provided their decisions. Therefore, the learners may have consulted their explicit knowledge - conscious knowledge - of prescriptive grammar concerning the syntactic rule under investigation. If this is typical, such confidence scales produce relatively unreliable results and do not reflect the learner's abstract syntactic knowledge.[9] In the end, "we need to be certain that these judgments are predicated on linguistic principles rather than on some other factors" (Goss, Ying-Hua, andLantolf, 1994, p. 264) so we are able to describe each learner's interlanguage competence.

Therefore, we can conclude that the rating scales (whether absolute or comparative) used in these and other similar studies, designed to investigate L2 grammatical competence, have necessarily influenced L2 learners' judgment behaviours - a conclusion which suggests that the results obtained from these studies are questionable, in the sense that they do not accurately reflect learners' unconscious grammatical knowledge.[10] However, this does not imply that the use of grammaticality judgements as an elicitation task should be abandoned in SLA research.[11] In fact, such a conclusion poses a serious challenge to linguists to find a better way of measuring acceptability – namely, an alternative rating scale that can overcome these methodological problems and pitfalls, so that the reliability of GJ tasks can be improved.

3. Measuring L2 learners' acceptability: an alternative rating scale

The discussion in the previous section about the major difficulties inherent in current acceptability measurement scales made it evident that a new rating scale that allows sharp lines to be drawn between the learner's certainty, doubt, and lack of knowledge reflected in his or her judgements need to be proposed. The following new rating scale can map the territory between these three possibilities that capture a learner's feelings towards any given sentence:

1. Bill wondered where Mary was going shopping.

---

[8] Although Gass (1994) did not provide interpretations, I think the sentences judged as +2 on Gass's rating scale may indicate that learners' grammatical knowledge included information allowing them to accept the structure of the sentences, but they avoided using them that way for different reasons. How do we differentiate these +2 from +1 interpretations, especially if we know that learners are found to "respond with greater certainty and accuracy to nondeviant strings than to deviant strings" (Hedgcock, 1993, p. 3), agreeing with Ellis (1991) and Sutter and Johnson (1990)? Additionally, things become more complicated when we attempt to provide different interpretations for the middle negative symbols -1 and -2.

[9] Within the UG framework, L2 researchers are interested in investigating learners' unconscious knowledge rather than their conscious knowledge (see White, 2003).

[10] Absolute rating scales can be binary (e.g., "grammatical" vs. "ungrammatical") or involve a third option (e.g., "don't know"), or involve different levels of (un)grammaticality (e.g., ''clearly grammatical'', ''possibly grammatical'', ''clearly ungrammatical'', and ''possibly ungrammatical''); whereas comparative rating scales require ranking sentences in comparison to one another on an acceptability rating scale (e.g., from ''sound perfect'' to 'sound horrible''). For more detailed discussion about the different the kinds of rating scales, refer to Schütze (1996).

[11] In fact, any data elicited by means of any other data elicitation technique (i.e., translation, story retelling, interview) must be influenced by test-related and/or subject-related factors such as age, mother tongue, intelligence, personality, motivation, language attitude, learning environment, teaching approach, the learner's test-taking strategies (i.e., guessing), nature of the linguistic input, topic of the task to be performed (see, Cook, 1990, 1996; Ellis, 1997; Littlewood, 1984).

☐ Clearly correct ☐ Clearly incorrect
☐ I don't know ☐ possibly incorrect
It should be: _____

As can be seen, this sentence needs to be judged on a 4-point scale: *clearly correct, clearly incorrect, possibly incorrect*, and *I don't know*.[12] Namely, L2 leaners need to judge whether the given sentence is acceptable by indicating clearly correct or clearly incorrect, but only if they are confident in their perception of the sentence. If the learner feels there is an error in the sentence but is not certain, he or she needs to judge the sentence as possibly incorrect. If the participant has no idea about the answer, he or she should choose don't know.

The words "It should be" followed by a blank space is provided, following Schütze's recommendation (1996), to prevent participants from making context less judgements. In a case where the response was clearly or possibly incorrect, the test-taker is asked to underline the perceived or possible errors in the sentence and provide a correction in the given space. This procedure will enhance the reliability of the gathered data because it allows the researcher to consider the correct unexpected responses, such when a respondent rejected a sentence based on a reason that was not actually related to its grammatical incorrectness. In this way, the possibilities of random and careless responses were minimised.[13] Besides helping to address these cases,[14] the double-check procedure (making judgements and providing corrections where required) assured the researcher that the data provided a relatively accurate reflection the learner's IL if the other performance factors were controlled for.

What is unique about the scale compared to others commonly used in grammaticality judgements (see section 2 above) is how the rating scale works and how the data obtained are marked and scored. Related to the former point, unlike in the scales previously used, only one doubtful category (*possibly incorrect*) was used in this scale (see the example above) to discriminate between the learner's certainty (both *clearly correct* and *clearly incorrect*) and lack of knowledge (*I don't know*) regarding a sentence. Each of the "possibly incorrect" options has the implied "possibly correct" meaning in addition to its literal meaning. This is because when a learner has doubt that a sentence is incorrect, he or she also has doubt that it is correct, regardless of the degree of doubt (little or great). Though they have the same embedded meanings, the reason behind the preference for including the category *possibly incorrect* in the test, rather than *possibly correct*, is to allow the researcher to ask the learner to correct the possible error he or she believes could render the sentence ungrammatical.

Related to the latter point, which considers the proposed method to be adopted in marking and scoring the obtained data collected by means of this rating scale, it is important to illustrate the fact that although the task required the participants to judge the test sentences on a four-point rating scale(*clearly correct*, *clearly incorrect*, *possibly incorrect,* and *I don't know*), the learners' responses to each sentence can be evaluated according to the following marking formula that covers all of their possible reactions to the sentence:[15]

1. clearly correct (CC)
2. clearly incorrect, and the right correction was provided (CIT)
3. clearly incorrect, but a wrong correction was provided (CIF)
4. clearly incorrect, but no correction was provided (CIN)
5. possibly incorrect, and the right correction was provided (PIT)

[12]Learners' intuitions in SLA research are reported interchangeably using a variety of terms – *grammatical*, *ungrammatical*, *acceptable*, *unacceptable*, *correct*, *incorrect*, *good*, *bad*, etc. Though there are theoretical distinctions drawn between these terms – see Birdsong (1989) – the rationale for choosing the terms *correct* and *incorrect* to report intuitions is the likelihood that L2 learners are used to such terms, especially those who are still taught via the use of grammar-translation methods. Richards, Platt, and Platt (1992) define the grammar-translation method as "a method of foreign language or second language teaching which makes use of translation and grammar study as the main teaching and learning activities" (p. 231).

[13] This is because unreliable (dishonest) participants often intentionally do not correct all the given ungrammatical sentences.

[14] See also the discussion below about the sentences marking criteria.

[15] Although the words *response* and *reaction* are often used interchangeably in the linguistics field, they will be used differently in this article. The word *response* will be used here only when referring to the four predetermined response options suggested to be given to test-takers in GJ task. The word *reaction* will be used to refer to possible various ways the participants react to the given responses – the nine possible reactions presented above.

6. possibly incorrect, but a wrong correction was provided (PIF)
7. possibly incorrect, but no correction was provided (PIN)
8. I don't know (DN)
9. Missing response (NA)[16]

This marking method, though it has not been used in published work, is suggested for the following reasons:

1. It allows the researcher to deal with the different reactions to the sentence that indirectly lead to the same concept/meaning, at least from the research goal's perspective. In other words, it enables the researcher, based on the correction provided, to regroup certain reactions that relatively share the same meanings, as follows:
   i. Any ungrammatical sentence rejected based on a reason that was not related to its grammatical incorrectness, whether it was judged to be either clearly incorrect or possibly incorrect, will be considered as if it had been judged clearly correct.
   ii. Any ungrammatical sentence judged to be possibly incorrect that was successfully rejected based on a reason related to its grammatical incorrectness will be counted as if it had been judged clearly incorrect.
   Note that because of the criterion in (ii), the *possibly incorrect* option will not be part of the analysis as a distinct category.
2. It gives the researcher the opportunity to handle some of the methodological problems commonly associated with participants' responses to judgements such as the missing responses, the incomplete responses as expected/required, and the*don't know* responses that can cause either significant data loss or biasing of the results, depending on the method of analysis adopted by the researcher to deal with such pitfalls. These problems can be addressed as follows:

i. Handling the missing reactions

Because it is not possible for the researcher to know whether a missing response is meaningful or not (e.g., whether the respondent left it out by mistake or omitted it intentionally because he or she did not want to answer it or had no idea about the answer), any missing responses are suggested to be excluded/removed from the statistical analysis.

ii. Handling the don't know responses

Given that such *don't know* responses can provide no information about the different/similar routes/stages through which learners pass in developing L2 grammatical knowledge, all such *don't know* responses to ungrammatical experimental sentences are suggested to be excluded/removed from the statistical analysis.

iii. Handling the incomplete responses as instructed

Because it is extremely difficult to know whether a sentence judged as either clearly incorrect or possibly incorrect was rejected based on reasons related to its grammatical incorrectness without providing a correction to that sentence, all such sentences are suggested to be excluded from the statistical analysis if the perceived or possible error in that sentence was not underlined or corrected.[17]

3. It allows the researcher to exclude the participants who do not perform the task as expected in a very methodical way. The method proposed to exclude such participants is based on meaningful information about the links (a) between the possible lack of seriousness in completing the task and providing no correction to some of the rejected sentences or providing no responses at all to a number of sentences and (b) between lack of sufficient

[16] I will consider in this article such missing values as a possible reaction to sentences. This is because respondents sometimes skip some questions intentionally, not by mistake, for various reasons (cf. Low, 1999; Dörnyei, 2003).

[17] Since it is commonly noted that there is a relationship between number of errors made and proficiency level and between type of errors and L2 learners' mother-tongue, it is sometimes possible for the researcher to figure out whether the uncorrected sentence was rejected for the reason that made it ungrammatical, especially with advanced L2 learners, who are expected to converge on native-like usage, and especially when comparing uncorrected sentences with other similar corrected sentences. However, this procedure could produce inconclusive results affected by the researcher's own opinion.

levels of proficiency in English to perform the task and providing too many don't know responses.[18] As a direct consequence of such links that could invalidate the task, the suggested criterion formulated to exclude such participants from the analysis is as follows:

> Any participant who has, say for example, 20% or more of his or her reactions to the experimental sentences excluded from the analysis based on the criteria stated in (2) above will not be included in the study.[19]

This suggested marking method requires, after marking the participants' reactions individually,[20] converting the performance on each sentence into a numerical score. Therefore, 1 point will be given for each reaction a participant made, regardless of its correctness. After that, these points for each of the nine possible reactions were calculated for every participant. Then, to calculate the percentage of the excluded responses for each participant in order to identify which participant(s) should be excluded from the statistical analysis according to the suggested 20% excluding criterion discussed above, the participants' scores of the four excluded reactions (1. Clearly incorrect – no correction was provided, 2. Possibly incorrect – no correction was provided, 3. I don't know, and 4. Missing response) of only the experimental sentences are added together. The following tables illustrate these marking, responses classification, and calculation processes:

Table 1: Methods suggested to mark the GJ test

Participant Number	Reactions to the Experimental Sentences						
	Sentence 1	…	Sentence 15	…	Sentence 34	…	Sentence 60
1	CIN	…	CIT	…	NA	…	CIT
2	CC	…	CIT	…	CIT	…	CIF
3	CIT	…	PIT	…	PIT	…	PIT
4	CC	…	PIN	…	DN	…	DN
5	NA	…	PIT	…	CC	…	DN
6	CC	…	CC	…	CIN	…	DN

Key:
CC: clearly correct
CIT: clearly incorrect – the right correction was provided
CIF: clearly incorrect – a wrong correction was provided
CIN: clearly incorrect – no correction was provided
PIT: possibly incorrect – the right correction was provided
PIF: possibly incorrect – a wrong correction was provided
PIN: possibly incorrect – no correction was provided
DN: I don't know
NA: missing response

Table 2: Classifying the possible responses to the judged sentences: included responses vs. excluded responses

Classification		Possible Responses to the Sentences								
		CC	CIT	PIT	CIF	PIF	CIN	PIN	DN	NA
Included Responses	Acceptance	✓			✓	✓				
	Rejection		✓	✓						
Excluded Responses							✓	✓	✓	✓

[18] If for any reason the proficiency test used failed to assign the learners to their correct levels, this procedure is likely to solve the problem - the threshold level required for participating in a particular study.

[19] This is just a suggested exclusion percentage. It is necessary to implement this percentage-based exclusion plan because "it is quite common to have a few missing values in every questionnaire" (Dörnyei, 2003, p. 106). Otherwise the researcher would end up losing a lot of data if any participant who did not complete all the task items as required were to be excluded from the analysis.

[20] Not according to the four predetermined response options given to the participants but according to scoring the possible reactions discussed above.

Table 3: Calculation processes suggested excluding reactions and participants in the GJ task

Participants	Excluded Responses		Exclusion
	Total	Percentage	
1	8	20 %	Included participant
2	5	13 %	Included participant
3	0	00 %	Included participant
4	20	53 %	Excluded participant
5	2	05 %	Included participant
6	24	63 %	Excluded participant

Tables 1, 2, and 3 above summarise, respectively, the procedure suggested to be used to mark the GJ task, the calculation processes suggested to exclude responses, and the results suggested to exclude participants. For example, based on the 20% excluding criterion, only participant's number 4 and 6 would be excluded from the statistical analysis, as the total number of their excluded reactions exceeded 20% of the total number of the expected reactions of the experimental sentences.

As for the remaining 5 possible reactions (CC, CIT, CIF, PIT, and PIF) to the ungrammatical sentences that will be included in the statistical analysis, they will be regrouped into only two categories (*acceptable English sentence* and *unacceptable English sentence*) based on the correctness/incorrectness of the correction provided to represent both the accepted and the rejected sentences. Table 4 explains the criteria suggested for regrouping them:

Table 4: Method suggested to be used in regrouping the included reactions: accepted vs. rejected English sentences

Participants	Included Responses to the Experimental Sentences				
	Total	Acceptable English Sentence (CC, CIF, PIF)		Unacceptable English Sentence (CIT, PIT)	
		Subtotal	%	Subtotal	%
1	22	4	18	18	82
2	25	9	36	26	64
3	28	0	0	28	100
5	26	19	73	7	27

This table shows that all the sentences rejected based on any reasons that were not related to their grammatical incorrectness will be considered as if they had been judged a clearly correct sentence. This is evident from the fact that three responses (CC, CIF, PIF), as they have the same meaning, are added together to form the category *acceptable English sentence*. On the other hand, this table shows that all the reactions to the ungrammatical sentences judged *clearly incorrect* or *possibly incorrect* for which right correction was also provided would be added together to form the category *unacceptable English sentence*.

In sum, this novel method of marking/scoring seems to offer two major advantages.[21] On one hand, by solving some of the methodological problems that can invalidate or at least bias the results of the GJ task, it ensures that for every participant, firm reliable conclusions can be drawn that reflect his or her individual knowledge in detecting the grammaticality/ungrammaticality of the given sentences in the contexts a researcher investigates. On the other hand, by making methodical changes in the data set through excluding and regrouping certain reactions, it appropriately prepares the data for statistical analysis by making it easier to handle, yet more reliable. A very important question arises from the above discussion – that is, whether this 4-point rating scale proposed in this article can produce reliable and valid data. This question can be answered scientifically only after conducting an empirical study using this scale. Nevertheless, the researcher's ability to exclude some participants from the analysis and distinguish which responses count as appropriate data may provide evidence that judgement data elicited using this rating scale are reliable in reflecting the learner's IL.

[21] Other marking methods used by researchers have been illustrated above in section 2 when describing the various rate scales used by them and the problems associated with them.

Note this is not to claim that this scale is free from problems, despite the fact that it has been argued that it attends to some of the inherent problems and limitations commonly associated with these kinds of rating scales used to measure L2 learners' acceptability. In a follow-up study, the scale's validity, reliability and limitations will be tested empirically using the test-retest procedure.

References

Adger, D. (2003). *CoreSyntax: A minimalist Approach*. Oxford, Oxford University Press.

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

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

Carroll, S. and J. Meisel. (1990). Universals and second language acquisition: Some comments on the state of the current theory. Studies *in Second Language Acquisition* 12(2): 201-208.

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

Chomsky, N. (1986). Knowledge of language: Its nature, origin, and use. New York, Praeger.

Cook. V. (1990). Timed comprehension of binding in advanced L2 learners of English. *Language Learning* 40(4): 544-599.

Cook, V. and M. Newson. (1996). *Chomsky's universal grammar.* Oxford, Blackwell Publishers Ltd.

Coppieters, R. (1987). Competence differences between native and near-native speakers. *Language* 63: 544-73.

Corder, S. P. (1973). The elicitation of interlanguage. In j. Svartvik (ed.), *Errata* (pp. 36-47). Stockholm: Rotobeckman.

Davies, W. and T. Kaplan. (1998). Native speaker vs. L2 learner grammaticality judgments. *Applied Linguistics*, 19, 2: 183-203.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, administration, and processing*. Mahwah, NJ, Lawrence Erlbaum.

Ellis, R. (1991). Grammaticality judgments and second language acquisition. *Studies in Second Language Acquisition 13*(2): 161–86.

Ellis, R. (1997) Second Language Acquisition Research and Language Teaching. Oxford: Oxford University Press.

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

Goss, N., Z. Ying-Hua and J. P. Lantolf. (1994). Two heads may be better than one: mental activity in L2 grammaticality judgments. In E. E. Tarone, S. M. Gass and A. D. Cohen (eds.), *Research methodology in SLA* (pp. 263-286). Hillsdale, NJ: Lawrence Erlbaum.

Haegeman, L. (1994). *Introduction to government and binding theory* (2nd ed.). Oxford, Blackwell.

Hedgcock, J. (1993). Well-formed vs. ill-formed strings in L2 metalingual tasks: specifying features of grammaticality judgements. *Second Language Research* 1: 1-21.

Keller, F. (1999). Grammaticality Judgments and Linguistic Methodology. *Journal of Logic, Language and Information* 8(1): 114-121.

Liceras, J. (1989). On some properties of the "pro-drop" parameter: Looking for missing subjects in non-native Spanish. In S.Gass and J. Schachter (eds.), *Linguistic Perspectives on Second language Acquisition* (pp. 109-133). Cambridge: Cambridge University Press.

Littlewood, W.T. (1984) Foreign and Second Language Learning: Language-Acquisition Research and Its Implications for the Classroom. Cambridge: Cambridge University Press.

Low, G. (1999). What respondents do with questionnaires: Accounting for incongruity and fluidity. *Applied Linguistics 20*: 503-33.

Pérez-Leroux, A. T. and W. Glass. (1997). OPC effects on the L2 acquisition of Spanish. In A. T. Pérez-Leroux and W. Glass (eds.), *Contemporary perspectives on the acquisition of Spanish. Vol. 1:Developing Grammar* (pp. 149-65). Somerville, MA: Cascadilla Press.

Richards, J. C., J. Platt and H. Platt. (1992). *Longman Dictionary of Language Teaching and Applied Linguistics*. London, Longman.

Schachter, J. (1989). Testing a proposed universal. In S. Gass and J. Schachter (eds.), *Linguistic perspectives on second language acquisition* (pp. 73-88). Cambridge: Cambridge University Press.

Schachter, J. (1990). On the issue of completeness in second language acquisition. *Second Language Research* 6: 93-124.

Schachter, J., and V. Yip. (1990). Grammaticality judgments: why does anyone object to subject extraction? *Studies in Second Language Acquisition* 12(4): 379-92.

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

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

Sutter, J.C. and C.J. Johnson. (1990). School-age children's metalinguistic awareness of grammaticality in verb form. *Journal of Speech and Hearing Research* 33: 84-95.

Tremblay, A. (2005). Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory. *Second Language Studies 24*: 129-167.

White, L. (1985). The pro-drop parameter in adult second language acquisition. *Language Learning* 35: 47-62.

White, L. (1986). Implications of parametric variation for adult second language acquisition. An investigation of the pro-drop parameter. In V. Cook (ed.), *Experimental approaches to second language acquisition* (pp. 55-72). Oxford: Pergamon Press.

White, L. (1992). Long and short verb movement in second language acquisition. *Canadian Journal of Linguistics* 37: 273-86.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge, Cambridge University Press.