Grammaticality Judgement Test: Does It Reliably Measure English Language Proficiency?

Bee Hoon Tan, Putra University of Malaysia, Malaysia

The European Conference on Language Learning 2014
Official Conference Proceedings

Abstract
Language assessment is an important activity in any language classroom. Out of the various tests or measurements of grammatical competence, one method known as grammaticality judgment is by far the most controversial albeit its advantages in gauging linguistic competence. Research studies on grammatical judgment tests (GJTs) are still getting contradictory research results since its introduction in second language research from the mid-70s (Rimmer, 2006). While productive language tests measure language use of learners (i.e. performance), GJTs are "a standard method of determining whether a construction is well-formed … where subjects make an intuitive pronouncement on the accuracy of form and structure in individual decontextualised sentences" (Rimmer, 2006, p.246). GJTs have been used to gauge linguistic competence of second language learners for more than three decades already, but the results differ. Several studies found GJTs reliable measures of learners' language competence (e.g. Leong et al., 2012; Rahimy & Moradkhani, 2012), while almost the same number found otherwise (e.g. Ellis, 2005; Tabatabaei & Dehghani, 2011). Therefore, this study aims to contribute empirical evidence to the field by administering a GJT to 100 ESL undergraduates. Comparison is made between the GJT scores and SPM English and MUET results to investigate the correlationship. A strong positive relationship among the three types of English proficiency measurement may indicate the reliability extent of the GJT as a measure of English language competence.

Keywords: grammar, grammaticality judgment, grammaticality judgment test, language competence, language performance

1.0 Introduction

Grammaticality judgment test (GJT) is one of the many ways to measure language proficiency and knowledge of grammar. It was introduced to second language research from the mid-70s. According to Rimmer (2006), GJTs are "a standard method of determining whether a construction is well-formed … where subjects make an intuitive pronouncement on the accuracy of form and structure in individual decontextualised sentences" (p.246).GJT is premised on the assumption that language proficiency comprises two types of language knowledge: receptive knowledge or language competence (i.e. knowing the grammar or metalinguistic awareness), and productive knowledge or language performance (i.e. using the language). Such tests are useful for the investigation of L2 learners' competence (abstract knowledge) not their performance (actual use of language in contexts) (Gass, 1994). Hence, GJT data reflect what the learners know and not what they do. In a GJT test, learners judge and decide if a given item, usually taken out of context, is grammatical or not.

Over the years, GJT has been used by researchers to collect data about specific grammatical features in testing hypotheses, and data collected by a GJT is said to be more representative of a learner's language competence than natural occurring data (Davies & Kaplan, 1998). It also allows the collection of negative evidence (ungrammatical samples) to compare with production problems such as slips and incomplete sentences (Schütze, 1996).

Despite the above mentioned usefulness of GJT, its application is riddled with controversies. Several studies found GJTs reliable measures of learners' language competence (e.g. Leong et al., 2012; Rahimy & Moradkhani, 2012), while almost the same number found otherwise (e.g. Ellis, 2005; Tabatabaei & Dehghani, 2011). Other than the reliability issues, it has been debated that certain item formatsof GJT are more reliable than others. The controversies related to GJT format can be related to, for example: selected versus constructed response, dichotomous versus multiple choice, ordinal versus Likert scale, and timed versus untimed test.

2.0 Purpose of the Study

The present study aims to investigate the reliability of a self-designed GJT in measuring the English language proficiency of 100 ESL undergraduates. The participants' performance scores on the GJT were compared to that of their SPM English and MUET results. The relationship was investigated by using Pearson correlation test. A high correlation index indicates high reliability between the two sets of test scores (GJT and SPM English; GJT and MUET) while a low correlation index indicates the low reliability in using the GJT to predict the performance on SPM English and MUET.

3.0 Research Methodology

The participants involved in the study were 100 undergraduates from two intact classes majoring in English language in a local public university. They were in the third year of their university studies.

The instrument was a self-designed GJT that was modeled after Gass (1994) and Salehi and Sanjareh (2013). The GJT comprises four sections with a total of 40

items. The first section on sentence grammaticality has 15 items with two response options each; the second section on correct word use has 5 items; the third section on gap-filling also has 15 items with three response format each; and the last section on sentence comprehension comprises 5 items. An example of each item format is as follows:

Section A: Sentence Grammar – Tick Correct or Incorrect to a given sentence.
Example: Most public buildings are air-conditioned and this means that any harmful tobacco smoke that are produced in one room will spread to other rooms through air-conditioning system.

Section B: Word Use - Select the sentence that uses the given word correctly.
Example: Whom

1. Most buy-out firms urgently need to return cash to investors, whom are impatient to see returns.
2. US authorities claim to have foiled the plot, whom, in the words of one of the alleged ringleaders.
3. He uses the world to interact with his son, with whom he has an estranged relationship.

Section C: Cloze test - Choose the right words in order to produce grammatical sentences.
Example: Symptoms of Alzheimer's usually develop slowly and get worse over time, becoming severe enough to interfere with daily tasks. Alzheimer's is a progressive disease, where dementia symptoms gradually _____over time.

a) worsens
b) worsening
c) worsen

Section D: Sentence comprehension - Select the option that has similar meaning with the given sentence.
Example: As for the Malaysian marine police, they have purchased high-speed boats, fast enough to catch pirate boats and durable enough to ram them if necessary.

1. The pirates own high-speed boats and they can escape from the Malaysian marine police very easily.
2. The government has allocated an amount of expenditure for the Malaysian marine police high-speed boats.
3. Once the Malaysian marine police detected the location of pirates, they chase and ram pirate boats if necessary using the high-speed boats they bought.

The GJT was conducted in class during a tutorial. The participants were told to write down their test start time and completion time on the test paper. Averagely they took between 15 to 25 minutes to complete the test. These data were collected for to find out if there is any relationship between test performance and

time spent on test. As such results are not within the scope of this paper, they will be reported in another paper.

4.0 Results and Discussion

After the scores of the GJT, SPM English and MUET were obtained, the scores were categorised according to three levels of language proficiency (high, intermediate, and low) so that they were compatible for comparison (see Table 1). For the GJT, the highest possible score was 40, and lowest possible score was zero. Hence, the scores that fell within the range of zero – 13 was set as the low level, the scores within the range of 14-26 was set as the intermediate level, and those within 27-40 belonged to the high level. For MUET, the lowest band was 1 and the highest band was 6. Therefore, the lowest proficiency level for MUET was Band 1 or 2, the intermediate was Band 3 or 4, and the highest level was Band 5 or 6. Subsequently for SPM English, the lowest possible grade was 1 and the highest was 9. Hence, the grades that fell within the range of 1-3 were counted as low level, grades 4 – 6 were moderate level and grades 7-9 were considered high level of English proficiency.

Table 1
Categorisation of language proficiency for GJT, MUET and SPM English

	GJT	MUET	SPM-English
High	27 - 40	5 - 6	7 - 9
Moderate	14 - 26	3-4	4 - 6
Low	0 - 13	1-2	1 - 3

The frequencies for each of the levels for the three tests were counted, and the results are as presented in Table 2. The results show that for the high proficiency level, about 80% of the participants scored the high proficiency level for SPM English, while about 30% was for GJT, and only about 8.5% was for MUET. On this basis, MUET would seem to be the most difficult English language for this group of participants. It should be noted that none of the participants was in the low proficiency group. The reason was because they were all majoring in English language, and the intake requirement was that they must scored at least an A in English and at least a Band 3 in MUET.

Table 2
Frequency of proficiency levels for GJT, MUET and SPM English

	GJT (n= 92)	MUET (n= 82)	SPM-English (n= 82)
High	30.42% (n=28)	8.54% (n=7)	80.49% (n=66)
Moderate	69.57% (n=64)	91.46% (n=75)	19.51% (n=16)
Low	00	00	00

With regards to whether there is any relationship between GJT with MUET and GJT with SPM English, the Pearson correlation index for GJT and MUET is 0.21 that indicates low level relationship (see Table 3). For GJT and SPM English, the Pearson correlation index is 0.03, indication no relationship between the two tests.

Table 3
Relationship between GJT and MUET, and GJT and SPM English (n=82)

	MUET	GJT
Pearson Correlation	1	.21
Sig. (2-tailed)		.06
Pearson Correlation	.21	1
Sig. (2-tailed)	.06	

	SPM	GJT
Pearson Correlation	1	.03
Sig. (2-tailed)		.80
Pearson Correlation	.03	1
Sig. (2-tailed)	.80	

Several reasons may be able to account for the non-relationship between GJT and MUET, and GJT and SPM English. Firstly the MUET and SPM English are meant to test mostly integrated skills of language production in various formats. Only a very small part of MUET and SPM English are designed to measure grammatical competence, therefore the comparison is incompatible. Secondly, the participants were in their final year of the undergraduate program. They sat for their SPM English and MUET over three years ago, and hence the related scores may not accurately represent their current proficiency level.

5.0 Conclusion

In hindsight, the current proficiency level of the participants should have been more accurately measured by a compatible productive test, and both the GJT and the productive tests should be set to test similar grammatical items. The GJT should have been more carefully designed to comprise only one type of test format so that confusion related to answering different test formats can be minimised.

Future research may also participants from across different disciplines and majoring in different area so that the data collected and the findings drawn could be more representative of the overall tertiary population of Malaysia. The present research study at best can serve as pilot research for improving the actual study.

References

Davies, W.D., & Kaplan, T.I. (1998).Native speaker vs. L2 learner grammaticality judgements. *Applied Linguistics, 19*, 183–203.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language: a psychometric study. *Studies in Second Language Acquisition*, *27*(2), 141-72.

Gass, S.M. (1994). The reliability of second-language grammaticality judgments. In E. Tarone, S.M. Gass, & A. Cohen (Eds), *Research methodology in second language acquisition* (pp.303–22). Hillsdale, NJ: Lawrence Erlbaum.

Leong, S.K., Tsung, L.T.H., Tse, S.K., Shum, M.S.K., & Ki, W.W. (2012). Grammaticality judgment of Chinese and English sentences by native speakers of alphasyllabary: a reaction time study. *International Journal of Bilingualism, 16*(4), 428-445.

Rahimy, R., & Moradkhani, N. (2012). The effect of using grammaticality judgement tasks on Iranian EFL learners' knowledge of grammatical patterns. *Asian Journal of Social Sciences and Humanities, 1*(2), 148-160.

Rimmer, W. (2006). Grammaticality judgment tests: Trial by error. *Journal of Language and Linguistics, 5*(2), 246–261.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3–13.

Salehi, M., & Sanjareh. H.B. (2013). The impact of response format on learners' test performance of grammaticality judgment tests. *Journal of Basic and Applied Scientific Research, 3*(2), 1335-1345

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

Tabatabaei, O., & Dehghani, M. (2011). Assessing the reliability of grammaticality judgment tests. *Procedia - Social and Behavioral Sciences, 31*(2012), 173 – 182. Retrieved from http://www.sciencedirect.com/science/article/pii/S1877042811029661

Contact e-mail: tanbh@upm.edu.my