Pertanika J. Soc. Sci. & Hum. 23 (S): 119 - 130 (2015)

SOCIAL SCIENCES & HUMANITIES

Journal homepage: http://www.pertanika.upm.edu.my/

Grammaticality Judgement Test: Do Item Formats Affect Test Performance?

Tan, B. H.* and Nor Izzati, M. N.

Department of English, Faculty of Modern Languages and Communication, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia

ABSTRACT

A grammaticality judgement test (GJT) is one of the many ways to measure language proficiency and knowledge of grammar. It was introduced to second language research in the mid 70s. GJT is premised on the assumption that being proficient in a language means having two types of language knowledge: receptive knowledge or language competence; and productive knowledge or language performance. GJT is meant to measure the former. In the 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 are said to be more representative of a learner's language competence than naturally occurring data. Collecting such data also allows the collection of negative evidence (ungrammatical samples) to be compared with production problems such as slips and incomplete sentences. Despite the usefulness of GJT, its application is riddled with controversies. Other than reliability issues, it has been debated that certain item formats are more reliable than others. Therefore, the present study seeks to determine if two different item formats correlate with the English language proficiency of 100 ESL undergraduates.

Keywords: Grammar, grammaticality judgment, grammaticality judgment test, item format, language competence, language performance

INTRODUCTION

A grammaticality judgement test (GJT) is one of many instruments used to measure language proficiency and knowledge of grammar. It was introduced to second language research in the mid 70s. According to Rimmer (2006), GJTs are "a

ARTICLE INFO

Article history:
Received: 8 July 2015
Accepted: 28 October 2015

E-mail addresses:
tanbh@upm.edu.my (Tan, B. H.)
norizzatimn@gmail.com (Nor Izzati, M. N.)
* Corresponding author

ISSN: 0128-7702 © Universiti Putra Malaysia Press

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 context) (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 naturally occurring data (Davies & Kaplan, 1998). It also allows the collection of negative evidence (ungrammatical samples) to be compared 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). Aside from reliability issues, it has been debated that certain item formats of 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 testing.

PURPOSE OF THE STUDY

The present study aims to determine if two different item formats correlate with the English language proficiency of 100 ESL undergraduates. The item formats tested were (1) sentence grammaticality: to judge if a given sentence is grammatical or ungrammatical by choosing from two options of *correct* or *incorrect*; and (2) gap-filling: to fill in blanks in a short paragraph by choosing from three options provided. The objectives of the research study were (1) to determine which of the two formats produced a higher mean score, and (2) to determine if there was any relationship between each of the item formats and the English language proficiency of the undergraduate subjects in the study as measured by the Malaysia University English Test (MUET).

REVIEW OF RELATED STUDIES

Related studies in the area of grammaticality judgment tests are reviewed below with respect to the issues of applications, reliability, item and response formats and new development.

Applications of Grammaticality Judgement Tests

Grammaticality judgement tests (GJT) are one of the established data-collection tools utilised to elicit information on grammatical competence, metalinguistic awareness and linguistic knowledge (Masny & D'Anglejan, 1985; Hsia, 1991; Andonova, *et al*, 2005). In L1 acquisition studies, GJT is conventionally used to determine if given structures are grammatical or ungrammatical in that language (Mandell, 1999), and in SLA research, they are employed to elicit data about the grammatical competence of students regarding a specific universal grammar (UG) principle or grammatical structure. This is "because it can provide crucial information about grammatical competence that elicited production tasks and naturalistic data collection cannot offer" (Tremblay, 2005, p.159).

Mackey and Gass (2005) described the GJT as a list of an approximately equal number of grammatical and ungrammatical sentences as stimuli on a target grammatical structure to which test-takers should respond as either correct or incorrect. In cases marked as incorrect, the correction should also be provided. They additionally recommended that the number of sentences not exceed 50, otherwise it may cause boredom. It is essential to include some fillers or distractors along with target sentences so that test-takers cannot easily speculate on the focus of the test.

The application of GJT, however, is not merely confined to grammatical competence. Hsia (1991), for instance, found that the ability to judge grammaticality is critical to reading for information and text interpretation. This was revealed through four tasks administered to 86 participants after reading a text. The first task embraced 10 true/false statements to measure their total comprehension; in this case, the test-takers were not allowed to look at the text again. The second task required the test-takers to reply to 10 multiple-choice comprehension questions, and the third was a GJT to test their ability to differentiate between deviant structures. The last one required them to fill in the missing parts of sentences based on their comprehension of the text; this task tapped into their metalinguistic competence of cohesion and discourse. The results displayed significant correlations between GJT and reading comprehension tasks and also dealing with cohesion and discourse, which was a metalinguistic type of task.

Tapping into metalinguistic awareness is another target of GJT; as Masny and D'Anglejan (1985) put it, "the operational definition of metalinguistic awareness is the grammaticality judgment test…it implies the ability to manipulate consciously various aspects of language knowledge" (p.179). In their study they explored the relationship between L2 learners' abilities to locate syntactically deviant structures and their cognitive and linguistic variables. To this end, variables such as cognitive style, intelligence, L2 aptitude, L2 proficiency, L1 reading and metalinguistic awareness in L2 were chosen. A GJT comprising three syntactic categories i.e. pronoun, relative clause and concord was constructed for the last variable. Among the results obtained in this study, statistical analyses showed that cloze tests, as a measure of integrative L2 proficiency, could reliably predict learners' ability to locate syntactic deviance. This suggests that "the ability to detect syntactic deviance can be considered a reliable correlate of second language competence" (Masny & D'Anglejan, 1985, p.186).

In addition to studies on metalinguistic awareness or knowledge, GJTs have been employed in research about individuals suffering from language impairment. Lely, Jones and Marshall (2011), for instance, employed a GJT to examine whether Grammatical-Specific Language Impairment children's errors in respect to wh-questions are caused by impairment in syntactic dependencies at the clause level or some other processes irrelevant to the syntactic system.

Eigsti and Bennetto (2009) utilised GJT to conduct research on children with autism to explore whether the way these children acquire the structures in their mother language differs from that of normal children, taking their developmental delay in the acquisition process into consideration. They argued that because GJT used in this study only necessitated judging heard sentences by the verbal response of yes/ no, it was a sensitively insightful device to evaluate structural knowledge of these participants.

Reliability of Grammaticality Judgement Tests

The reliability of the grammaticality judgment test (GJT) in second language acquisition research has been a matter of concern for many researchers. Ellis (1991) is one of the first who employed a test-retest research design in his study to address the reliability of grammaticality judgments in second language acquisition. His study had two phases with a one week interval between them. In both phases of the experiment, advanced ESL Chinese students were asked to make judgments about sentences involving dative alternation in English. In the second phase, some of the participants were also asked to perform a think-aloud task. Based on the considerable inconsistency observed in his participants' grammaticality judgements, Ellis suggested that "learners' judgments can be inconsistent, and therefore unreliable, when they are unsure" (Ellis, 1991, p.181). He maintained that beginners are not suitable subjects for examining the reliability of GJT because their judgment data are not validated by data from other types of tasks (e.g. oral production).

In another study, Mandell (1999) compared data from GJTs with data from Dehydrated Sentence Tests (DSTs) (a slash-sentence test that is commonly used in the L2 classroom to examine L2 learners' knowledge about word order) in order to investigate the reliability of GJTs. Data were collected from three levels (second, fourth and sixth semesters) of adult L2 learners of Spanish. The results from the comparison of the two tests indicated that "a definite relationship existed between the standard GJT and the DST" and "the grammaticality judgments of L2 learners, although indeterminate, were consistent." (Mandell, 1999, p.93). Therefore, Mandell concluded that GJTs were reliable measures of L2 learners' linguistic competence.

Unlike Ellis' (1991) study, the study conducted by Tabatabaei and Dehghani (2012) involved advanced learners who were selected using the Oxford Placement Test (OPT). The researchers implemented GJT with a test-retest design that was divided into two categories: timed GJT, where the learners needed to answer the test in a given period, and delayed GJT where they were given flexible answering time. Participants were asked to make judgments about 34 sentences included in a computerised GJT. The grammatical structure chosen for this study was verb complements. The results of the test-retest analysis and internal consistency reliability revealed that the GJT used in this study had a low level of reliability. Moreover, the analysis of response patterns showed that participants were not stable in their judgments and also, they were reluctant to use the "not sure" response when they were uncertain. Therefore, their judgements did not exactly reflect their grammatical knowledge. Finally, the relationship between timed GJT and delayed GJT was weak, which indicated that participants may have used different types of knowledge under different test administration conditions. The results of this study suggest that the GJT used in this study was not a reliable measure of EFL learners' knowledge about verb complements, and researchers should use this kind of test with caution.

Schütze (1996) identified the measurement scale, instructions and subject-related factors as the linguistic and non-linguistic factors that might influence judgement behaviour and, hence, engender instability and unreliability. The researcher suggested that the measurement scale (including nominal, ordinal and interval scales) used for judgement elicitation is crucial as it determines what type of data is obtained and which mathematical (statistical) operations can be carried out on the data. The instructions used in judgement elicitation have considerable influence on the outcomes of experiments. In most experiments, the speakers who function as subjects are naive and, hence, likely to be unfamiliar with the linguistic concepts that they are supposed to apply in rating the stimuli. If no definitions for grammaticality are provided, each subject will use his or her own interpretation of these concepts, and the resulting data are likely to be very noisy .

Item and Response Format of Grammaticality Judgement Tests

Hohensinn and Kubinger (2011) stated that there are two classes of item formats. First is the constructed response formats, which are also called "open-ended" or sometimes "free response formats," that demand the examinee or test-taker to create and write down the solution ranging from single words up to a few sentences. For this format, the examinee has to generate his or her ideas on a particular theme and compose a longer text passage. The second format is the multiple-choice format. It requires the examinee (test-taker) to choose the right answer(s) from several given answer options. Conventional multiple-choice formats offer a single correct answer option and one or two to seven distractors. Other multiple-choice formats contain more than a single solution that the examinee has to mark. This latter multiple-mark format has already been recommended by Cronbach (1941) and Pomplun and Omar (1997) as a feasible alternative to conventional multiple-choice items.

Haladyna, Downing and Rodriguez (2002) divided the multiple-choice (MC) format into seven categories: conventional MC, alternate-choice, matching, true-false, multiple true-false (MTF), context-dependent items, including the item set and complex MC. The researchers indicated that conventional MC, true-false and matching are three formats that have scored 100% for frequency of citation in 27 textbooks on educational testing, and the results of 27 research studies and reviews published since 1990.

Shizuka *et al.* (2006) investigated the effects of three- and four-option items on test performance within the context of an L2 English reading test used as a university entrance exam in Japan. They changed an original four-option reading test to a three-option test by discarding the least-chosen option from a previous administration of the test. One hundred and ninety-two Japanese English-language learners who had not taken the original test took the revised test. Just like the outcomes from educational measurement research, their results indicated that the average item facility and average item discrimination between the four-option-item test and the three-option-item test were not significantly different. Also, test reliability was not significantly different across test formats. Furthermore, in their analysis of distractors, they found that the average number of actual plausible distractors was less than two, regardless of the number of options the items had. Thus, the researchers claimed that items with three options are optimal, considering three- and four-option items had relatively equal item facility and item discrimination.

Currie and Chiramanee (2010) conducted a study in the context of L2 testing that investigated how multiple-choice items differ from open-ended items in measuring L2 English grammar. Relevant to research investigating the optional number of options in multiple-choice items, they included three versions of the multiple-choice test in their investigation: three-, four- and five-option versions. They found three-option items were easier for the learners in their study (L2 English learners in Thailand), but there were no significant differences in item facility between the four- and five-option items. They noted that multiple-choice testing is widespread in ESL and EFL contexts worldwide; thus, it is important, they wrote, that researchers come to understand how L2 learning outcomes are shaped by the type of L2 tests learners take (2010, p.488).

The effect of test response formats was investigated by Salaehi and Sanjareh (2013). Their study compared two pairs of test items: multiple-choice GJT (MCGJT) versus dichotomous GJT (DGJT) and ordinal GJT (OGJT) versus Likert GJT (LGJT). The results showed that subjects performed better in DGJT and LGJT. The researchers did not discuss the outcomes very much. They only highlighted how distinct response formats can influence subjects' performance. They even supported their findings with the study conducted by Rodriguez (2005), who concluded that the number of options in multiple-choice tests affects reliability, item difficulty and item discrimination. Analysing 27 studies that dealt with different response formats of MCQ, he asserted that three-option multiple choice tests are optimal. Having investigated a varied range of reductions i.e. reduction of options from 5 to 4, 3 and 2, and also the decrease of 4-option items to 3- and 2-option items, he found that 3-option items are optimal since shifting from 4- to 3-option items raises reliability slightly by .02 and item discrimination by .03.

Development of New Grammaticality Judgement Tests

One of the most commonly used Grammaticality Judgment Tests (GJT) is multiple-choice questions, better known as MCQs. The standard multiple-choice format has remained relatively unchanged for nearly 100 years, even over the past 25 years when multiple-choice tests became computerised. A psychologist, psychometrician and recognised luminary in the measurement industry, David Foster, is credited with introducing computerised adaptive testing (CAT) and simulation-based performance testing as part of Novell's IT pioneering certification programme in the early 1990s.

The newly developed CAT is called discrete-option multiple-choice or DOMC. The DOMC item format uses the basic elements of the traditional multiple-choice or Trad-MC (Foster & Miller, 2009), format stem and answer options. The essential difference lies in randomly presenting the options one at a time on the screen and asking the test-taker to decide if the option that appears is the correct one or not. The item is considered to be completed when the test-taker demonstrates that she or he has answered the item correctly or incorrectly.

An example of a DOMC item using the content of a mathematical question that was given by Foster and Miller (2009) is shown below. In this example, the answer option shown (number 29) is the correct answer and was randomly selected for presentation on the screen.

Q. Is this number a prime number?

29

Yes No

With the DOMC format, there is only one way for a test-taker to answer an item correctly, which is to choose Yes when the correct option is displayed. There are two ways for a test-taker to answer a question incorrectly: (a) Choose Yes when an incorrect option (or distractor) is displayed, or (b) choose No when the correct option is displayed. The item continues and provides another answer option if the test-taker chooses No when an incorrect option is displayed. In the study conducted by Foster and Miller (2009), five answer options were used. This means the test-taker needed to provide either a ‘yes’ or ‘no’ response to the remaining four answer options before s/he could proceed to the next question.

Foster stated that the computerised version of the multiple-choice format mainly emphasises the aspect of security. In DOMC, the answer options are displayed randomly, which indicates that each test-taker will get different sequences of answer options. For instance, in order to answer a question, test-taker A will have to answer either Yes or No to option 1 followed by the other four options i.e. option 2, option 3, option 4 and option 5. Test-taker B may encounter the options in this sequence: option 5, option 3, option 1, option 2 and option 4.

Since each answer option is presented separately in DOMC, unlike the traditional multiple-choice (Trad-MC) which exposes all the options at once, the test items are unlikely to be memorised or captured through technology and shared with others. Foster and Miller (2014) stated that Trad-MC items are prone to being stolen and later re-used. Braindump sites, that is, websites where stolen test content is sold, proliferate. Test items are often discussed openly on Web forums and in chatrooms. Moreover, in psychometric parlance, DOMC is a way to prevent the occurrence of construct irrelevant variance (CIV) elements, which are test-taking skills (test-wiseness) and cheating. The prevention of CIV elements is acknowledged to be helpful in neutralising the unfair stigma that has been associated with Trad-MC.

Despite the improvements offered by DOMC, there are still inconveniences that need to be considered. Because not all of the answer options are presented in DOMC or because they are only presented one at a time, either of these situations may result in shorter or longer amounts of time to complete each item. This is because each person has a unique style i.e. the intellectual functioning as well as personality type that pertain to a person as an individual that makes the individual different from others (Brown, 2007). Brain hemisphere dominance and reflectivity and impulsivity are two qualities of styles. Style is seen as an important issue to be taken into account when dealing with DOMC. A resolution needs to be figured out so that the items format is universal and applicable to all groups of test-taker.

In order to meet needs that include computers and maintenance, the DOMC software, which is apparently costly, and training for teachers could require a large budget to cover expenses. Plus, getting used to the new DOMC assessment system could, of course, demand an extension of time since the Malaysian education system has been engaged with ‘paper and pencil’ examination systems since its inception.

RESEARCH METHODOLOGY

The participants involved in the study were 100 undergraduates from two on-going 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 modelled after Gass (1994) and Salehi and Sanjareh (2013). The GJT comprised two sections. The first section on sentence grammaticality had 15 items with two response options each and the second section on gap-filling also had 15 items with a three-response option format each. An example of each item format is as follows:

Section A:
Sentence Grammaticality

Example 1: The increasing number of abandoned newborn babies is a serious social concern as a large number of mothers who dump their babies are underage and unmarried.

Correct [ ]
Incorrect [ ]

Section B: Gap-Filling

Example 1: Alzheimer's is a progressive disease, where dementia symptoms gradually ____(1)____ over time.

1. a) worsens [ ]
   b) worsening [ ]
   c) worsen [ ]

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. On average, they took between 15 to 25 minutes to complete the test. These data were collected to determine if there was any relationship between test performance and time spent on the test. As such results are not within the scope of this paper, they will be reported in another paper.

When the test papers were marked, it was found that eight participants left some items unanswered. To ensure better reliability of the results, the scores of the incomplete tests were not included in the calculation.

RESULTS AND DISCUSSION

Overall, the participants' performance on Section B Gap-Filling with three-response options was better than their performance on Section A Sentence Grammaticality with two-response options. The results show that the mean score for Section B was 9.8 compared to 8.43 for Section A (Table 1).

TABLE 1
Descriptive Statistics for Two- and Three-Option Formats (n = 92)

	Two-Option	Three-Option
Mean	**8.43**	**9.80**
Median	8.00	10.00
Mode	7.00	11.00
Std. Deviation	1.97	2.50
Minimum	3.00	3.00
Maximum	12.00	14.00

The higher mean score in the third column of Table 1 shows that the gap-filling three-option response format was less difficult than the sentence grammaticality two-option response format. This is supported by the literature, which seems to suggest that a three-option response format is more reliable than a four- or five-option response format (see, for example, Rodriguez, 2005). According to Rodriguez (2005), three options are the optimal response format since shifting from 4- to

3-option items raises reliability slightly by .02 and item discrimination by .03. Studies conducted by Shizuka, Takeguchi, Yashima and Yoshizawa (2006) as well as Currie and Chiramanee (2010) also obtained the same finding.

Although logically the two-option format has a higher percentage of getting a correct answer by chance alone (50% compared to about 33% for three options, 25% for four options and 20% for five options), according to Fagan (2001), a two-option response format has lower reliability and less discrimination than response formats, possibility of bias with regards to test-wiseness, response-style and guessing, and is often only suitable for factual information.

With regards to whether there is any relationship between each of the two test formats and the participants' English language proficiency based on MUET, the results (see Table 2) show that there is no relationship between test format and MUET. For this computation, the sample size dropped to 83 because only 83 out of the 92 participants stated their MUET scores.

TABLE 2
Correlation between Test Format and MUET (n=83)

	Pearson Correlation Index
MUET	
Sentence Grammaticality Two-Option	0.164
Gap-Filling Three-Option	0.238

Several reasons may account for the lack of a significant relationship between the test formats and MUET in the sample. Firstly, MUET is meant to test mostly integrated skills of language production in various formats. Only a very small part of MUET is designed to measure grammatical competence. Therefore, the comparison is incompatible. Secondly, the participants were in their final year of the undergraduate programme. They sat for MUET over three years previously, and hence, the MUET score may not have accurately represented their proficiency level at the time of the study.

CONCLUSION

In hindsight, the GJT should have been more carefully designed. The research meant to investigate whether GJT test formats affect test performance and whether there is any relationship between each test format and MUET test scores. Although no relationship was found in the latter, there is a positive answer in that participants performed better in the gap-filling three-option format. However, it should be pointed out that in the present study, two aspects of the GJT were tested: the item format and the response format. Hence, the results cannot be confidently attributed to the test format alone.

Future research could standardise the response format to strictly focus on the item format. For example, the response options should be of the same number and the discrete grammatical items being tested should also be the same to produce more confident results and findings.

REFERENCES

Andonova, E., Janyan, A., Stoyanova, K., Raycheva, M., & Kostadinova, T. (2005). *Grammaticality judgment of article use in aphasic speakers of Bulgarian*. Unpublished manuscript.

Brown, H. D. (2007). *Principles of language learning and teaching.* New York, NY: Pearson Education Inc.

Cook, V. (2003). *The innateness of a universal grammar principle in L2 users of English*. *IRAL*. Retrieved on Jan 16, 2014 from http:// homepage.ntlworld.com/ vivian.c/ Writings/ Papers/SD&UG.htm

Currie, M., & Chiramanee, T. (2010). The effect of the multiple-choice item format on the measurement of knowledge of language structure. *Language Testing, 27*(4), 471-491.

Davies, W. D., & Kaplan, T. I. (1998). Native speaker vs. L2 learner grammaticality judgements. *Applied Linguistics, 19*(2)*,* 183-203.

Eigsti, I. M., & Bennetto, L. (2009). Grammaticality judgments in autism: Deviance or delay. *Journal of Child Language, 36*(5), 999-1021.

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

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

Fagan, J. C. (2001). Selecting test item types to evaluate library skills. *Research Strategies, 18*(2), 121-132.

Foster, D., & Miller, H. L. (2009). A new format for multiple-choice testing: Discrete-option multiple-choice. Results from early studies. *Psychology Science Quarterly, 51*(4), 355-369.

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.303-322). Hillsdale, NJ: Lawrence Erlbaum.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 5*(3), 309-334.

Hohensinn, C., & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*(4), 732-746.

Hsia, S. (1991). Grammaticality judgments, paraphrase and reading comprehension: Evidence from European, Latin American, Japanese and Korean ESL learners. *Hong Kong Journals Online, 3,* 81-95.

Kamisah Omar, & Nurul Aini Bakar. (2012). Educational computer games for Malaysian classrooms: Issues and challenges. *Asian Social Science, 8*(1), 75-84.

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.

Mandell, P. B. (1999). On the reliability of grammaticality judgement tests in second language acquisition research. *Second Language Research, 15*(1), 73-99.

MacDonald, J. L. (2000). Grammaticality judgments in a second language: Influences of age of acquisition and native language. *Applied Psycholinguistics, 21*(03), 395-423.

Mackey, A., & Gass, S. M. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Erlbaum.

Masny, D., & D'Anglejan, A. (1985). Language, cognition, and second language grammaticality judgments. *Journal of Psycholinguistic Research, 14*(2), 175-197.

Pomplun, M., & Omar, M. H. (1997). Multiple-mark items: An alternative objective item format. *Educational and Psychological Measurement, 57*(6), 949-962.

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

Rice, J. W. (2007). New media resistance: Barriers to implementation of computer video games in the classroom. *Journal of Educational Multimedia and Hypermedia, 16*(3), 249-261.

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, IL: University of Chicago Press.

Shizuka, T., Takeuchi, O., Yashima, T., & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing, 23*(1), 35-57.

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

Tremblay, A. (2005). Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory. *Second Language Studies, 24*(1), 129-167.