J. Basic. Appl. Sci. Res., 3(2)1335-1345, 2013

ISSN 2090-4304
Journal of Basic and Applied Scientific Research
www.textroad.com

The Impact of Response Format on Learners' Test Performance of Grammaticality Judgment Tests

Mohammad Salehi*, Hemaseh Bagheri Sanjareh

Assistant Professor Languages and Linguistics Center, Sharif University of Technology

ABSTRACT

The grammaticality judgment (GJ) test constructed by Gass (1994) was selected as the instrument which was manipulated in terms of response format while the original items remained intact. Therefore, in the first phase of the study, a multiple-choice GJ test (MCGJT) was developed to be compared with the traditional form, i.e., the dichotomous type (DGJT) in terms of participants' test performance. The two tests were administered to a group of 110 engineering-major BS students at Sharif University of Technology. In the second phase, GJ tests were developed in ordinal (OGJT) and likert (LGJT) scales. 49 engineering- BS students at the same university participated. The analysis unveiled the effect of response format on test performance.
KEY TERMS: Reliability, Validity, Grammaticality judgment, Dichotomous, Multiple-choice, Ordinal scale, Likert scale.

1. INTRODUCTION

Grammaticality Judgment (GJ) tests are one of the established data-collection tools utilized to elicit information on grammatical competence, metalinguistic awareness and linguistic knowledge (e.g., Hsia, 1991; Masny & D' Anglejan, 1984; Andonova, Janyan, Stoyanova, Raycheva & Kostadinova, 2005). In L1 acquisition studies, GJ tests are conventionally used to examine if the 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).

Considering the importance of these tests in L1 and SLA studies, it is crucial to investigate the effect of response format of GJ tests on the individuals' test performance. This is due to the fact that no research has ever been undertaken regarding the influence response format might leave on the results of GJ tests. Since expected response, according to Bachman (1990), can be determined through test design and be elicited via proper instructions, task specification and input, it is part of the test method. Moreover, it is noteworthy that test method effect is considered as one of the systematic measurement errors affecting test scores.

Thus, this study aims to explore the impact of response format on the respondents' performance on these tests. So far, the conventional forms of GJ tests have had either a dichotomous or a gradient approach to grammatical competence, in which the latter is mainly adopted through the application of a Likert scale (Gass & Ard, 1984; Schachter & Yip, 1990). In the present research, ordinal and likert scales, as well as multiple-choice and the traditional dichotomous formats will be incorporated in responses. Meticulous examination of such an impact will subsequently make a contribution to obtaining more reliable and valid results providing insightful information for test developers, teachers and other stakeholders.

The Research Questions

1. Does response format (i.e., multiple-choice versus dichotomous) affect test performance?
2. Does response format (i.e., likert versus ordinal) affect test performance?

2. REVIEW OF THE LITERATURE

2.1 Response Format

Features of test methods, alternatively termed facets, affect test performance which is partly owing to the fact that the individuals' characteristics, i.e., cognitive and affective styles, interact with the aspects of the test methods (Bachman, 1990). Hence, he asserted that performance on language tests is the outcome of the interaction between a testee's language ability and other variables not targeted by the research such as cognitive and affective characteristics and features of the test method. Chappelle (1988), for instance, probed the influence of a cognitive factor such as field independence as a probable source of variance on cloze, dictation, multiple-choice and essay tests which were administered to 224 participants of native and non-native. As one of his findings, field-independence, both among natives and non-natives, had only a high and positive correlation with multiple-choice tests lending support to the interaction between test format and cognitive styles of learners.

***Corresponding Author:** Mohammad Salehi, Assistant Professor Languages and Linguistics Center, Sharif University of Technology. Tel.: +98 021 66164732 , Email address: m_salehi@sharif.ir

Among diverse characteristics of test method, item stimulus and response format are distinguishing components of a test (Cohen, 1980, as cited in Bachman, 1990). Thus, the response formats [test method] selected for testing language ability may itself exert an influence on the student's score, and since the impacts of the response format tend to be unpredictable, it can potentially be a source of construct-irrelevant variance (Alderson, Clapham, & Wall, 1995).

David (2007) explicates some reasons for this issue, first of which is that certain constructs may be restricted or prevented by item format to be incorporated in the test. Item format may also induce interference with the construct; consequently yielding contaminated scores which are not purely reflective of the construct or language ability in question. Increasing the chance of coverage for other components of the construct, and leading the test-takers to think in specific ways not intended by the researchers are among the other underlying reasons. It is also worthy of note that, according to him, each of these effects of format might manifest itself at varying levels of competency with differing degrees.

Some researches have been undertaken concerning this issue among which investigations of constructed-response and multiple-choice formats have received most attention. Tsagari (1994), for instance, compared the effects of constructed-response items with multiple-choice items on tapping reading ability. 57 individuals were presented with the two content-equivalent passages with differing formats along with a checklist of test-taking strategies and retrospective questionnaires pertaining to more general reading strategies. The findings indicated that multiple-choice items demanded distinctively different response strategies, and that these two test types tapped different constructs. The results of this study additionally implied that method effects can be a source of threat to the validity of scores and results of a test in that they might measure constructs differing from those the research seeks to.

In the same vein, Kobayashi (2002) addressed the impact of two factors of text organization and response format on test-takers' scores of reading comprehension. The assumption was that since tests are constructed to evaluate the learners' language abilities, they should be as least affected and intervened as possible by other variables such as response format and text organization. Thus, the instrument comprised texts of four rhetorical organizations along with three types of response formats i.e., cloze, open-ended questions and summary writing. Significant differences in test performance were found across the text types and response formats suggesting that different response formats gauge different aspects of reading comprehension ability.

Rodriguez (2003) undertook a comprehensive review of previous research into construct equivalence of the same two formats, i.e., constructed-response and multiple-choice formats. Out of 67 studies, 29 identified studies reported 56 correlations between items in both formats. A meta-analysis of disattenuated correlations revealed that tests with stem equivalence yielded a synthesized coefficient of near to one. This is indicative of the tests being congeneric, whereas items having stems with differing wordings or formats were comparatively far less correlated.

Likewise, in another study by Currie and Chiramanee (2010), the effect of multiple-choice format, juxtaposed with a constructed-response test, was investigated on the measurement of knowledge of language structure. To this end, a test of English structure in constructed-response format and, afterwards, in three multiple-choice formats containing 3-, 4- and 5- choices were administered to one hundred fifty-two university students. Although the scores of the two tests were found to be highly correlated pointing to the same construct they measured, a direct comparison of answers to the items in the two tests revealed that only 26% of them were the same. For them, this discrepancy denotes that what multiple-choice tapped plainly relied on the item format. The researchers, therefore, concluded that despite all the benefits they offer like practicality and objectivity while employing multiple-choice instruments, one needs to be circumspect about the risk of contaminating the construct by the influence of item format.

Shohamy (1984) conducted a study exploring the impact of two different test methods, multiple-choice tests and open-ended questions, on measuring reading comprehension besides the use of L1 and L2. The results revealed that different test methods, producing varied levels of difficulty, leave differential influences on participants. In her study, open-ended questions were found to be more demanding particularly for participants with lower proficiency. This, as Weir (2005) asserts, lies in the fact that our choice of format will immensely influence the cognitive processes the task involves.

This study attempts to contribute to this line of research by filling the existing void, i.e., lack of studies on the effect of GJ response formats on the individuals' performance.

2.2 Grammaticality Judgment Tests and Their Applications

To infer from the abstract area of grammaticality, linguists may occasionally need to resort to a form of forced elicitation to further probe into a particular grammatical structure (Mackey & Gass, 2005). This approach is, in some cases, essential in that employing avoidance strategy, individuals shy away from producing those structures they are not confident about or they have not fully internalized. Besides, it might be considerably time-consuming to await the natural production of the intended structures on the part of the participants during which the inevitably developing nature of interlanguage should also be taken into account.

Mackey and Gass (2005) describe the GJ test as a list of approximately equal number of grammatical and ungrammatical sentences as stimuli on a target grammatical structure to which test-takers should respond as correct and incorrect. In case marked as incorrect, the correction should also be provided. They additionally

recommend 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.

Grammaticality judgment is an elicitation tool employed extensively to obtain concrete information about the abstract nature of UG and grammatical competence which is of main interest in this perspective study (Tremblay, 2005; Cook, 2003). In this respect, Tremblay (2005) asserts that the use of GJ tests in challenging linguistic theories is necessitated by the valuable and useful information it can yield of which the common production tasks and naturalistic data collection methods are incapable.

The application of GJ tests, however, is not merely confined to grammatical competence. Hsia (1991), for instance, found out 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 ten true/false statements to measure their total comprehension without having the permission to look at the text again. The second task required the test-takers to reply to ten multiple-choice comprehension questions, and the third one was a GJ test to test their ability of differentiating deviant structures. For the last one, they were asked to fill the missed parts of sentences based on their comprehension of the text tapping individuals' metalinguistic competence of cohesion and discourse. The results displayed significant correlations between GJ and reading comprehension tasks and also last one, dealing with cohesion and discourse which was a metalinguistic type of task.

Tapping metalinguistic awareness is another target of GJ test as Masny and D' Anglejan 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' (1985, p. 179). In their study they explored the relationship between L2 learners' ability to locate the 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 GJ test, comprised of three syntactic categories, i.e., pronoun, relative clause and concord, was constructed for the last variable. Among the other results obtained in this study, statistical analyses showed that cloze tests, as a measure of integrative L2 proficiency could reliably predict the learners' ability to locate the 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 the same vein, as a dissertation for her Ph. D degree, Renou (1998) probed the influence of grammaticality judgments on the relationship between metalinguistic awareness and L2 proficiency. The test-takers were sixty-four French L2 learners with the proficiency level of high-intermediate to advanced at University of Ottawa. Presented in both modalities of written and oral, the GJ test was comprised of 9 grammatically correct and 21 grammatically incorrect French sentences. In this research, metalinguistic awareness was operationally defined by means of the scores obtained by GJ test. The results confirmed the first hypothesis that metalinguistic awareness was closely related to L2 proficiency.

In addition to studies on metalinguistic awareness or knowledge, GJ tests have been employed in research about individuals suffering from aphasia, autism and broadly speaking language impairments and disorders. Lely, Jones and Marshall (2011), for instance, employed a GJ test 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. The reason for the choice of the GJ test as the instrument lies in the fact that, according to Tyler (1992, as cited in Lely, Jones & Marshall, 2011), it is specifically effective while examining testee's with aphasia since it provides the researchers with the opportunity to differentiate impairments in the stored syntactic knowledge from those that take place later in the processing procedures.

Eigsti and Bennetto (2009) utilized GJ tests to conduct research on children with autism to check 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 argue that because GJ test used in this study only necessitated judging the heard sentences by the verbal response of yes/no, it was a sensitively insightful device to evaluate the structural knowledge in these participants.

GJ tests have been utilized in several studies to assess the morphosyntactic knowledge. Fledge, Yeni-Komshian and Liu (1999), for instance, administered a 144 item- GJ test, to 240 native speakers of Korean, with various age on arrival (AOA) in the U.S, to examine the critical period hypothesis for SLA. The knowledge of morphosyntax was evaluated through the GJ test which was presented aurally. The test-takers were also required to repeat 21 English sentences containing a variety of consonants and vowels as a measure for foreign accent. The results showed that as AOA increased, the foreign accents became stronger and the scores on GJ test gradually decreased.

Likewise, Khamis-Dakwar, Froud and Gordon (2011), used GJ to explore the morphosyntactic knowledge of Modern Standard Arabic (MSA) and Palestinian Colloquial Arabic (PCA) on one hundred-twenty Arabic-speaking children. Two GJ tests each containing forty items were administered to children. The results displayed that when two constructions were similar in the two language varieties, children performed better on those items.

3. METHODOLOGY

3.1 Participants

Regarding the first research question, a total of 110 students participated in the study majoring in various engineering fields (e.g., Computer, Electronic, IT and etc.) at Sharif University of Technology (SUT). They were all freshmen and within the age range of 18 or 19. The participants came from five different classes, the homogeneity of whom was ensured through their mid-term scores. Prior to test administration, each of the constructed tests was piloted on adult English learners at Kish language institute. 25 participants took part in the piloting phase of multiple-choice grammaticality judgment tests. To ascertain the compatibility of participants in the sample and the piloting phase in terms of the proficiency level, an attempt was made to select the sample from upper-intermediate, which was close to the overall proficiency of the main population based on several observation sessions.

With respect to the second research question, totally, 44 B.S-engineering students at Sharif University of Technology took part in the three phases of the research and 49 participated in the first two phases, five participants missing the last part. They were all freshmen and in the age range of 18 or 19 coming from three different general English classes. Their homogeneity was established by means of their mid-term scores.

3.2 Instrumentation

3.2.1 The Original Instrument

The instrument of this study, a grammaticality judgment test (GJ), was originally developed by Gass (1994). The instrument comprises 24 items and 7 distractors added by the researchers 'so that participants in a study cannot easily guess what the study is about' (Mackey & Gass, 2005, p. 51). Guessing what the test focuses on, according to them, can be a threat to the internal validity in that the results are affected by the factors other than the ones the study aims at. The target grammar of this test is based on relative clause positions on the accessibility hierarchy, initially proposed by Keenan and Comrie (1977 & 1979). The hierarchy manifests the extent to which the relativization of NP positions can be accessible.

SU > DO > IO > OBL > GEN > OCOMP

The above abbreviations on the hierarchy respectively stand for subject, direct object, indirect object, oblique case, genitive, and object of comparison. It reflects the concept that subjective relative clauses are more accessible than direct objective clauses and the latter is more accessible than indirect objective ones and so forth. Hence, the test embodies 6 subsets: SU, DO, IO, OPREP, GEN, and OCOMP. There are 4 sentences for each subset, two of which are grammatically incorrect, and two are correct.

3.2.2 The Modified Instruments

The original test, i.e., the dichotomous GJ test (DGJT, see Appendix A) was transformed into three other different versions in terms of response format. The multiple-choice grammaticality judgment test (MCGJT henceforth, see Appendix B), ordinal grammaticality judgment test (OGJT, Appendix C) and likert grammaticality judgment test (LGJT, Appendix D). In all of these tests, the items constructed by Gass (1994) were kept intact, the only distinguishing feature being their response formats. Each item in the GJ tests is either grammatically correct or incorrect. Regarding the MCGJT, items are followed by two headings of correct and incorrect under each of which three choices are provided. As explained in the instruction of the test, the respondents are required to select either correct or incorrect options. Concerning choices under the incorrect option, three words of the sentence are presented and for correct option, three sentences which are, in fact, the entailments of the sentence in question provided. Therefore, if, for instance, the item is identified as correct by the respondent, he/she selects the correct heading and subsequently chooses one of the three sentences which best entails the item. It is only in this case that the test-taker receives the full point of that question. Selecting the correct heading with a wrong choice would not buy them any point. Likewise, recognized it as incorrect, the respondent needs to select the wrong choice and provide the correct form afterwards. The selection of grammatically incorrect part and provision of the correct form would end in full point of that question. If any of these steps were done wrongly, the respondent would not obtain any points

In the OGJT, each item embraces three choices being only distinct in terms of the grammatical correctness of the relative clauses. In other words, one choice is grammatically 'correct', one is 'incorrect' and the other is termed 'most incorrect'. 'Incorrect' and 'most incorrect' differ with respect to the number of grammatical mistakes; with the former containing one deviant form and the latter more than one. Hence, the test-taker is required to rank the choices of each item based on their degree of correctness writing the number of the choices under the corresponding column. In this way, for each question is scored either, 0, 1 or 3.

Concerning the LGJT, the choices are presented on a five-point likert scale ranging from definitely correct to definitely incorrect. The testee needs to select the most appropriate choice in his/her viewpoint.

3.3 Justification for the Test Selection

Inappropriate selection of test content, namely the mismatch between items and objectives, may lead to a source of test invalidity (Henning, 1987). Thus, it is of paramount importance for a test to closely correspond with the objectives of the study along with the target population especially in terms of their level of proficiency. One of

the main reasons underlying the choice of the GJ test by Gass (1994) as the instrument of the current study was that the target grammar of the test in question was in accordance with the proficiency level of the participants based on several observations and content of their English course book. Majority of standard GJ tests developed by scholars of this area were concerned with UG-based principles and parameters, e.g., pro-drop parameter, dative alteration and subjacency, which would not present much challenge for the participants of the current study, being too easy for their proficiency level and, therefore, resulting in lack of response validity. Moreover, since the study Gass (1994) conducted on her developed GJ test, dealing with the reliability of L2 GJs, was to some extent related to this research, the researchers deemed it right to select this test as the instrument.

3.4 The Mid-term Exam

The mid-term exam of Sharif University of Technology was employed to establish the homogeneity of the participants coming from five different classes. The test has been developed by the "Languages and Linguistic Faculty" members and, therefore, enjoys the construct validity crucial to any developed test via the expert judgments. This is well explicated by Alderson et al. (1995), who assert that expert judgment is required for both content validity and construct validity to ensure the correspondence of the test with its underlying theory. The test comprised 40 items of vocabulary and grammar being administered in two versions which only differed in the order of the items. This was undertaken to attenuate the possibility of cheating which is a threat to internal validity (Henning, 1987). The reliability of both versions was 0.87 Cronbach coefficient alphas.

3.5 Data Collection Procedure

3.5.1 Piloting

As noted above all the constructed tests were initially piloted on a sample of Persian EFL female adult learners at Kish language Institute. This largely benefited the test construction and administration procedures in a number of ways. Having first administered the MCGJT test on a sample of 11 students at upper-intermediate level, which was a corresponding level to the target group's proficiency based on the several observations, the researchers received insightful feedbacks. For instance, based on subsequent retrospective interviews with some students and their responses to the MCGJT, ambiguous and flawed choices were identified and singled out for revision; meanwhile, the appropriate timing for the test was checked. The participants' comments and the process of responding the test, i.e., 50 minutes, led the researchers to reduce the number of choices for each item from 4 to 3. This modification resulted in reduced test time, i.e., 30 minutes which subsequently resulted in more practicality of the test, since it became less time-consuming and tiring to the respondents. To further ascertain the accuracy of choices and modifications applied, the revised MCGJT was piloted for the second time on a different group of 13 students who were in another class but at the same proficiency level. The second administration, however, yielded successful results and, henceforth, no need for additional revisions was seen.

A piloting process was also performed on the OGJT due to its novel and probably unfamiliar response format for test-takers. The test was administered to 12 adult EFL learners in the same institute, who took the test in almost 25 minutes. Since no pitfalls were noticed in this phase, the test was deemed suitable for the target sample.

3.5.2 The Main Data

For the first research question, the data were collected through five intact General English classes at Sharif University of Technology in the first semester. Prior to administration of tests, the students were required to be cognizant of the time not exceeding it so that they mainly relied on their intuitions and also could not return to change their responses. There was a month interval between the MCGJT and the DGJT administration during which the researchers made sure that the participants were not exposed to any instructions pertinent to the target structures.

Regarding the second research question, the data were obtained from three intact General English classes at the same university in the second semester. The same care about time constraints was exercised, and the test-takers were furnished with precise instruction regarding the format due to the fact that the OGJT was completely new and unfamiliar to them. A two-month interval was considered between the administration of the OGJT and the LGJT.

3.6 The Design

The design of the study is ex post facto which, as Ary, Jacobs and Razavieh (1996) state, is undertaken after variation in the desired variable has already been established in the natural line of events. This design is called causal comparative as well in that it attempts to examine cause-and-effect relationships between independent and dependent variables. This design is, however, applied to situations where randomization of participants and manipulation of variables as well as application of a treatment, being among prominent features of experimental research, are not permitted. For Hatch and Lazaraton (1991), an ex post facto design is the most prevalent design type in applied linguistics in that it permits us to probe what is happening rather than what caused this.

Data Analysis

To probe the impact of response format, i.e., multiple-choice vs. dichotomous, on the test-takers' performance, a paired-samples t-test was run comparing the mean scores and variances on the two tests. To

investigate the effect of these two response formats, i.e., the OGJT and the LGJT, on test performance, a paired samples t-test and chi-square were also employed.

4. RESULTS

4.1 Descriptive Statistics

Descriptive statistics was employed to pave the grounds for running analysis of variance. The results are presented in Tables 1 and 2.

Table 1. ***Descriptives Statistics for the Mid-term Scores of the Five Classes***

Group	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
1	25	29.8400	6.18250	1.23650	18.00 38.00	
2	21	26.3810	6.20867	1.35484	13.00	35.00
3	22	30.4091	5.86076	1. 24952	17.00	39.00
4	22	27.0000	7.82548	1.66840	8.00	37.00
5	20	29.9500	7.27270	1.62622	15.00	40.00
Total	**110**	**28.7455**	**6.77615**	**.64608**	**8.00**	**40.00**

As displayed in Table 1, the means of the classes are 29.8, 26.3, 30.4, 27 and 29.9 for class 1, 2, 3, 4 and 5 respectively. This speaks to negligible differences among the classes.

With regard to second research question, the descriptive statistics of the data obtained from three classes are as follows:

Table 2. ***Descriptive Statistics for the Mid-term Scores of the Three Classes***

Class	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
1	18	24.3333	5.21311	1.22874	7.00	29.00
2	23	24.0435	3.63666	.75830	17.00	30.00
3	25	22.2800	6.05888	1.21178	2.00	29.00
Total	**66**	**23.4545**	**5.09957**	**.62771**	**2.00**	**30.00**

As can be observed in Table 2, the means of the classes are 24.3, 24 and 22.2, for class 1, 2, 3 respectively. This reflects the insignificant differences among the classes.

4.2 One-way ANOVA

A one-way ANOVA was run to provide robust evidence for homogeneity of the classes. As Table 3 reveals, regarding the first and second research questions, the F-observed value is 1.69 (P = .15 > .05) which is lower than the critical value of 2.45 at 4 and 105 degrees of freedom. Since the F-observed value is lower than its critical value, it can be concluded that there are no significant differences among the five groups of participants.

Table 3. ***One-way ANOVA for Mean Differences among the Five Classes***

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	304.292	4	76.073	1.699	.156
Within Groups	4700.581	105	44.767		
Total	5004.873	109			

Regarding the second research question, as displayed in Table 4, the F-observed value is 1.08 (P= .34 >.05) which is lower than the critical value of 3.14 at 2 and 63 degrees of freedom. Since the F-observed value is lower than its critical value, it can be concluded that there are no significant differences among the three groups of participants.

Table 4. ***One-way ANOVA for Mean Differences among the Three Classes***

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	56.367	2	28.184	1.087	.344
Within Groups	1633.997	63	25.936		
Total	1690.364	65			

4.3 First Phase

4.3.1 Paired-samples T-test

To investigate whether response format affects test performance, a comparison of the means of the participants on the DGJT and the MCGJT was performed through the paired-samples T-test.

Table 5. ***Paired Samples Statistics for the DGJT and the MCGJT***

	Mean	N	Std. Deviation	Std. Error Mean
DGJT	19.2545	110	4.28018	.40810
MCGJT	15.3727	110	5.13105	.48923

The mean scores of the two tests, as displayed in Table 5, disclose the major difference between the two performances.

Table 6. ***Paired Samples Test for the DGJT and the MCGJT***

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
DGJT- MCGJT	3.88182	3.62137	.34528	3.19748	4.56616	11.242	109	.000

Table 6, reveals that there is statistically a significant difference between the DGJT and the MCGJT in terms of student performances (t (109) = 11.24, P = .00 < .05). This is suggestive of the fact that the two stem-equivalents tests of differing formats did not measure exactly the same construct.

Based on Graph 1, test takers in all the five groups performed better on the DGJT. This is reflective of the influence of response format on the test performance.

4.4 Second Phase

4.4.1 Chi-square

Regarding the second research question, a chi-square was run to explore the probable influence of these two response formats on participants' test performance. This was achieved through counting the frequency of correct and incorrect responses in each test. Thus, the extent to which response format played roles in the existent inconsistency could be revealed.

Table 7. ***Cross-tabulation between Response Format and Response***

			R		Total
			Correct	Incorrect	
RF	**Likert**	Count	2658	870	3528
		Expected Count	2417.0	1111.0	3528.0
		% within RF	75.3%	24.7%	100.0%
		% within R	55.0%	39.2%	50.0%
		% of Total	37.7%	12.3%	50.0%
	Ordinal	Count	2176	1352	3528
		Expected Count	2417.0	1111.0	3528.0
		% within RF	61.7%	38.3%	100.0%
		% within R	45.0%	60.8%	50.0%
		% of Total	30.8%	19.2%	50.0%
Total		Count	4834	2222	7056
		Expected Count	4834.0	2222.0	7056.0
		% within RF	68.5%	31.5%	100.0%
		% within R	100.0%	100.0%	100.0%
		% of Total	68.5%	31.5%	100.0%

Table 8. ***Chi-square Tests***

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	1.526E2[a]	1	.000		
Continuity Correction[b]	151.984	1	.000		
Likelihood Ratio	153.532	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	152.595	1	.000		
N of Valid Cases[b]	7056				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 1111.00.

b. Computed only for a 2x2 table

As can be observed in Table 8, the correlation between the two variables was significant, X^2 (1, N=49) = .00, p <.05. This shows that response format has significantly affected the test performance.

4.4.2 Paired-samples T-test

To further examine the relation between response format and test performance and also the superiority of one response format over the other, a paired-samples t-test was conducted.

Table 9. ***Paired Samples Statistics for the LGJT and the OGJT***

	Mean	N	Std. Deviation	Std. Error Mean
LGJT	54.2449	49	12.29080	1.75583
OGJT	44.6735	49	13.79521	1.97074

Table 10. ***Paired Samples Test for the LGJT and the OGJT***

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
LGJT- OGJT	9.57143	11.49819	1.64260	6.26876	12.87409	5.827	48	.000

The results arrived at via the chi-square are supported by the paired-samples t-test as well in that, as Table 10 reveals, there is statistically a significant difference between the LGJT and the OGJT performances ($t(49) = 5.82$, $P = .00 < .05$). Additionally, based on Table 9, the participants obtained better scores from the LGJT as opposed to the OGJT. This indicates that factors other than the target construct, e.g., test-taking strategies, are being measured in these stem-equivalent tests in differing formats.

According to Graph 2, individuals in the three groups performed better on the likert scale than the ordinal one and this speaks to the effect of response format on the respondents' test performance.

5. DISCUSSION AND CONCLUSION

Drawing upon the pertinent literature, it has been held by many scholars (e.g., Bachman, 1990; Alderson et al., 1995; Weir, 2005), that response format as one of the characteristics of test method affects the test performance. However, no research has ever examined such impact on GJ tests and they are commonly used in two formats of dichotomous and likert. The statistical analyses conducted through the T-test and the Chi-square for the two research questions lent support to this argument. Thus, as a source of systematic measurement error, response format can affect reliability and consequently validity of a test.

The second research question addressed the effect of two different response formats, i.e., the LGJT and the OGJT, on the test performance. These formats were at odds with those of the first research question, i.e., the MCGJT and the DGJT. However, the results obtained for the former were in conformity with the latter. Additionally, it is also probable that different response formats exert influences to differing degrees. This might have been closely associated with the extent test-takers perceive these formats to be marked to them since 'the consistency speaks of a level of cognitive behavior related not only to item content and test taker knowledge/ability but also to the format in which the item is presented to the test taker'(Currie & Chiramanee, 2010, p. 487).

As noted in the introduction section, the primary purpose of this study was to explore the impact of response format on the test performance of GJ tests. Afterwards, the statistical procedures were run on each test. In conformity with the pertinent literature, it seems that response format can induce variations in test performance and affect it.

The findings encourage language testing professionals and researchers to reflect upon the problem of response format as a systematic measurement error on test-takers' test performance. Therefore, it is advisable to control this factor prior to the application of these tests, particularly in the most conventional formats, i.e., dichotomous and likert.

6. Implications and suggestions for further research

While transforming the original GJ test into tests with differing formats, the researchers came up with ideas for GJ test construction. First and foremost, based on the rigorous study by Rodriguez (2005), the number of options in multiple-choice tests affects reliability, item difficulty and item discrimination. Analyzing 27 studies in this respect, he asserted that three-option multiple choice tests are optimal. It is of immense importance; therefore, to consider this issue while developing an MCGJT. This is due to the fact that number of options has the potentiality to be a source of systematic error for these tests causing item difficulty, for instance, to function as an uncontrolled independent factor.

Second, item writing for the OGJT should be carried out with circumspection in that any minor changes in the structure, which are irrelevant to the target grammar being investigated, can make the sentences more difficult. In this vein, item difficulty would function as an uncontrolled factor. For instance, inclusion of passive grammar as a marked structure in the options would result in far more challenging and difficult items. Nonetheless, incorporating any grammatical variations associated with the target structure is recommended. In this study, for example, articles were included since their use is closely related to relative clauses.

The following insights can be of immense avail to future researchers:

1. Despite the fact that during administration phase of the study, care was exercised regarding timing, since timing is of utmost importance in GJ tests, the researchers noticed that some participants finishing early returned

and changed their responses. Thus, conducting a computer-based method of this study would shed more light on this area controlling the time participants are allowed to spend on each test and subsequently highlighting the factor of intuitive judgment about grammaticality or ungrammaticality of the sentences.

2. The impact the number of options in multiple-choice tests exerts on reliability, as only one of the dependent variables of this specific study, is well highlighted by Rodriguez (2005) and therein lies another potential research area. Having investigated varied range of reductions, i.e., reduction of options from 5 to 4, 3, 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. Hence, to contribute to this line of research, the influence of number of options in the MCGJTs on the reliability and validity of these tests could fill the existing void.

3. Because age, literacy, education and idiolect are factors whose effect has been studied on GJ tests (Tremblay, 2005). Therefore, another procession of research can explore the probable interaction between different response formats of GJ tests and test-takers' language abilities, i.e., language proficiency level, individual characteristics such as age, L1 background, language, education (Bachman & Palmer, 1996).

7. REFERENCES

1. Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

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

3. Ary, D., Jacobs, L. C., & Razavieh, A. (1996). *Introduction to research in education.* Florida: Harcourt Brace College.

4. Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

5. Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

6. Cook, V. (2003). The innateness of a universal grammar principle in L2 users of English. *IRAL*. Retrieved August 12, 2011, from

http://homepage.ntlworld.com/vivian.c/Writings/Papers/SD&UG.htm

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

8. David, G. (2007). Investigating the performance of alternative types of grammar items. *Language Testing, 24*(1), 65–97.

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

10. Fledge, J. E., Yeni-Komshian, G. H., & Liu, S. (1999). Age constraints on second- language acquisition. *Journal of Memory and Language, 41*, 78-104.

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

12. Gass, S. M., & Ard, J. (1987). L2 acquisition and the ontology of language universals. In W. Rutherford (Ed.), *Second language acquisition and language universals* (pp. 33-68). Amsterdam : John Benjamins.

13. Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguistics.* New York: Newbury House.

14. Henning, G. (1987). *A guide to language testing*. Cambridge, Mass: Newbury House.

15. 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. Retrieved August 15, 2011, from

http://sunzi.lib.hku.hk/hkjo/article.jsp?book=10&issue=100004

16. Keenan, E. L., & Comrie, B. (1977). Noun phrase accessibility and universal grammar. *Linguistic Inquiry, 8*(1), 63-99.

17. Keenan, E. L., & Comrie, B. (1979). Noun phrase accessibility revisited. *Language, 55*(3), 649-664.

18. Kobayashi, M. (2002). Method effects on reading comprehension test performance: Test organization and response format. *Language Testing, 19*(2), pp. 193-220.

19. Khamis-Dakwar, R., Froud, K., & Gordon, P. (2012). Acquiring diglossia: Mutual influences of formal and colloquial Arabic on children's grammaticality judgments. *Journal of Child Language, 39, 61-89.*

20. Lely, H. K. J., Jones, M., & Marshall, C. R. (2011). Who did buzz someone? Grammaticality judgment of wh-questions in typically developing children and children with grammatical-SLI. *Lingua, 121,* 408-422.

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

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

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

24. Renou, J. M. (1998). *Effect of grammaticality judgments on the relationship between metalinguistic awareness and second language proficiency.* Unpublished doctoral dissertation, University of Ottawa, Canada.

25. Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163-184.

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

27. Schachter, J., & Yip, V. (1990). Why does anyone object to subject extraction? *Studies in Second Language Acquisition, 12*(4), 379-392.

28. Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147-170.

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

30. Tsagari, C. (1994). *Method effect on testing reading comprehension: How far can we go?* Unpublished MA thesis, University of Lancaster, UK.

31. Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Basingstoke: Palgrave Macmillan.

Appendix A. The Dichotomous Grammaticality Judgment Test

……**1**. I saw the man who crossed the street……..
……**2.** This is the woman whom I am taller than. ……

Appendix B. The Multiple-choice Grammaticality Judgment Test

1. I saw the man who crossed the street.

Incorrect	**Correct**
1. The man……	1. I looked at the man while crossing the street.
2. Who……	2. The man who crosses the street knew me.
3. The street……	3. I saw a man and he crossed the street.

2. That is the woman whom I am taller than.

Incorrect	**Correct**
1. Than……	1. That woman is taller than me.
2. The woman……	2. I am taller than the woman.

Appendix C. The Ordinal Grammaticality Judgment Test

1.

1. I saw the man whom crossed the street.
2. I saw the man who crossed the street.
3. I saw the man whom he crossed the street.

Correct	Incorrect	Most incorrect

2.

1. That is the woman which I am taller than her.
2. That is the woman whom I am taller than.
3. That is the woman who I am taller than her.

Correct	Incorrect	Most incorrect

Appendix D. The Likert Grammaticality Judgment Test

Definitely grammatical Probably grammatical Unsure Probably ungrammatical Definitely ungrammatical

1. I saw the man who crossed the street.

1 2 3 4 5

2. That is the woman whom I am taller than.

1 2 3 4 5

Appendix E. The Post-test Questionnaire of Validity

Please choose one of the items below as a reason for your choice.

I chose this order in the first test because:

a) I was 100% sure that the order I wrote was correct.

b) I was 50-99% sure that the order I wrote was correct.

c) I was less than 50% sure that the answer was correct.

d) The format of the test confused me.

e) I did not read the sentences properly.

f) I did not know the answer so I guessed.

If not any of the above, please write your reason for the answer you gave:

c) I was less that 50% sure that the answer was correct

d) The format of the test confused me.

e) I did not read the sentences properly.

f) I had learned the answer since taking the first test.

If not any of the above, please write your reason for the answer you gave here:
