TELL, Vol. 8, No. 1, Spring & Summer 2014, pp.93.122

The impact of response format on validating grammaticality judgment tests

Mohammad Salehi[1]
Assistant Professor, Sharif University of Technology, Iran
Hemaseh Bagheri Sanjareh
M.A. in TEFL, Sharif University of Technology, Iran

Received on November 19, 2012
Accepted on July 2, 2013

Abstract

The present research aims to examine the influence of response format on the reliability and validity of grammaticality judgment tests. To this end, the grammaticality judgment test developed by Gass (1994) was selected as the instrument which was manipulated in terms of the response format while the original items remained intact. Thus, in the first phase of the study, a multiple-choice GJ test was constructed to be compared with the traditional form, i.e., the dichotomous type. The two tests were administered to a group of 110 students. With regard to validity which was arrived at by means of internal correlations proposed by Alderson, Clapham and Wall (1995). In the second phase, GJ tests were developed in ordinal and likert scales to which 49 students responded. The analysis unveiled the effect of response format on reliability. Concerning validity, which was examined via a post-test questionnaire, response format

[1] Corresponding author: Sharif University of Technology
Email address: m_salehi@sharif.ir

was selected as the second reason behind inconsistency in responses from participants' view.

Keywords: reliability, validity, grammaticality judgment, scales

1. Introduction

Grammaticality Judgment (GJ) tests are one of the most prevailing data-collection tools employed to elicit information on grammatical competence, metalinguistic awareness and linguistic knowledge (e.g., Andonova, Janyan, Stoyanova, Raycheva & Kostadinova, 2005; Hsia, 1991; Masny & D'Anglejan, 1984;). GJ tests are conventionally used in L1 acquisition studies to examine if the given structures are grammatical or ungrammatical in that language (Mandell, 1999). Yet, in SLA research, these tests are employed to elicit data about the grammatical competence of students regarding a specific universal grammar 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). However, Due to lack of consensus among researchers on the consistency of judgments and the genuine nature of these tests, i.e., what they actually measure, their reliability and validity are still widely disputed.

To overcome the deficiencies of GJ tests, a myriad of approaches have been adopted. Yet, no research has ever been undertaken regarding the impact response format might leave on the reliability and validity of GJ tests and, therefore, there is a void to be filled by exploring the effect response format can potentially leave on the reliability and validity of these 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 the reliability and consequently validity of test scores.

Thus, this study aims to explore the influence of response format on the reliability and validity of these tests, which will also reveal some facts about the superiority of each format. 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. 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.

2. Review of Literature

2.1 Response Format

Attributes of test methods, alternatively termed facets, influence test performance which is in part due to the fact that individual characteristics, i.e., cognitive and affective styles, of test-takers interact with the features of the test methods (Bachman, 1990). Thus, he stated that performance on language tests is an 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. Chapelle (1988), for instance, examined the effect 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 her findings, field-independence, both among natives and non-natives, only correlated highly and positively with multiple-choice tests lending support to the interaction between test format and cognitive styles of learners.

Among varied characteristics of test methods, item stimulus and response formats 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 research has 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 respondents 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 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 respondents' scores of reading comprehension. The assumption was that since tests are developed to measure 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.

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 those 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.

2.2 Reliability and Validity of GJ Tests

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 the

current 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.

Despite the widespread utilization of this test in linguistic theories and ample reliance placed on it in conducting research, the reliability, and validity of GJ tests are still the controversial subject of debate due to some observed inconsistencies. It is crucial to ensure the reliability of data obtained via this test otherwise the data will be fallacious exemplars of grammatical structures leading the researchers to draw erroneous conclusions (Gass, 1994).

Mandell (1999) found consistent and cross-sectional correlations between data of GJ tests and a dehydrated- sentence test about the syntax of V-movement in Spanish with participants from three levels of second, fourth and sixth semesters. The observed correlations were found not only within one level of test-takers but also across the three different levels of learners. He, therefore, concluded that GJ tests are reliable measures of L2 competence in this respect. The findings of his study were in conformity with those of Gass (1994) who investigated the reliability of L2 GJ tests. Her developed test comprised the six different types of relative clauses and the judgments were compared across ESL learners of China, Korea and Japan. The reliability was calculated by means of test-retest method with a one-week interval. Despite some variation in performance between the two administrations, the participants' performance was totally reliable.

Ellis (1991) examined whether L2 judgments were similar to L1 judgments made about dative alteration in English by 21 adult advanced Chinese ESL learners. To this end, he employed a 40-item GJ test along with some think-aloud protocols with eight of the participants. He concluded that 'the study does indicate that grammaticality judgments are potentially unreliable' (p. 181). In his opinion, the performance on these tests is affected by factors such as the subject's stage of development, the items being tested and the

nature of the test itself. This is also well supported by Gass (1994) who stated that reliability of GJ is inextricable from indeterminate responses which can be attributed to the incomplete stage of learning.

Besides some noted inconsistencies in judgments, the validity of GJ tests is open to debate as well, that is, the extent to which they are accurately reflective of grammatical competence. To address the validity issue of grammaticality judgments and investigate the assumption that GJ in SLA entails the same underlying activities in L1, Davies and Kaplan (1998) employed two techniques. They compared the think-aloud protocols on GJ tests administered to the participants, who were fourth- term learners of French taking the test in their L1, i.e., English and a subset of them taking twelve sentences of the same test in French, which was their L2, as well. The findings were classified based on the strategies observed. The comparison of the L1 and L2 GJ test results revealed that the strategies employed on L2 GJ tests outnumbered those of L1 and differed in type as well.

Tremblay (2005) came up with some reasons to approach this issue. First and foremost, similar to other data collecting tools, GJ tests are affected by extragrammatical (performance) variables among which Sorace (1996, pp. 377-378) names: 'parsing strategies, context and mode of presentation, pragmatic considerations, and linguistic training and mental or introspective state'. Other opponents of GJ tests as valid measures, name L1 influence and test-taking strategies learners were taught as factors affecting the judgments (Davies & Kaplan, 1998). In the same vein, some scholars contend that GJ tasks cannot provide direct access to linguistic competence (e.g., White, 2003) 'because grammaticality is not open to direct introspection' (Tremblay, 2005, p. 134). Another factor influencing GJ negatively is the lack of control over methodological aspects of this test such as the materials, procedures and analysis and interpretation of results.

3. Purpose of the Study

The research questions were:

1. Does response format (i.e., multiple-choice versus dichotomous) affect reliability and validity of grammaticality judgment tests?
2. Does response format (i.e., Likert versus ordinal) influence reliability and validity of grammaticality judgment tests?

4. Method

4.1 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. 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.

4.2 Instrumentation

4.2.1 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.

4.2.2 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, 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.

4.3 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 respondents 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 researcher deemed it right to select this test as the instrument.

4.4 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 alpha.

4.5 Data Collection Procedure

4.5.1 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.

4.5.2 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.

4.6 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.

4.7 Data Analysis

Reliability is obtained through Cronbach's alpha for all response formats. Concerning the validity, with regard to the first research question, as Alderson et al. (1995) assert, internal correlations are employed as one of the means of investigating construct validity. According to Alderson et al. (1995), these correlations are gained by correlating the different test subsets. The logic lies in the fact that the test subsets are to measure different factors and, as a result, have contributory roles to the total linguistic ability which underlies the test. Thus, we presumably expect the components to yield rather low correlations, i.e., 0.3-0.5. Internal correlations, as one of the

approaches to internal validity, are adopted by some other researchers as well such as Kyzlinkova (2007), and Yujie and Wenxia (2007).

With respect to the second research question, validity was probed via a post-test questionnaire (see Appendix E), most items of which were adopted from Currie and Chiramanee (2010). This questionnaire aimed at exploring the test-takers' reasons behind their inconsistent responses for an item in both tests. In doing so, the researchers furnished the participants with the two tests and marked that specific item to which they had replied completely differently. The researchers added one item of their own and asked the participants to select their corresponding reason for each of the tests. They were, moreover, asked to write their reasons, in a blank box provided, if not among the listed items. The frequency of the item selected for each test was computed and analysed.

5. Results

5.1 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: *Descriptive statistics for the mid-term scores of the five classes*

Group	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
1	25	29.84	6.18	1.23	18.00	38.00
2	21	26.38	6.20	1.35	13.00	35.00
3	22	30.40	5.86	1. 24	17.00	39.00
4	22	27.00	7.82	1.66	8.00	37.00
5	20	29.95	7.27	1.62	15.00	40.00

Total	110	28.74	6.77	.646	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 shown 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.

Table 2: *Descriptive statistics for the mid-term scores of the three classes*

Class	N	Mean	Std. Deviation	Std. Error	Minimum	Maximum
1	18	24.33	5.21	1.22	7.00	29.00
2	23	24.04	3.63	.75	17.00	30.00
3	25	22.28	6.05	1.21	2.00	29.00
Total	66	23.45	5.09	.62	2.00	30.00

5.2. 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	304.29	4	76.07	1.69	.156
Within Groups	4700.58	105	44.76		
Total	5004.87	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	56.36	2	28.18	1.08	.34
Within Groups	1633.99	63	25.93		
Total	1690.36	65			

5.3 Reliability Estimation

Cronbach's coefficient alpha was employed being one of the most prevalent statistics of internal consistency (Pallant, 2005).

Table 5: *Reliability statistics for the MCGJT*

Cronbach's Alpha	N of Items
.84	24

Table 6: *Reliability statistics for the DGJT*

Cronbach's Alpha	N of Items
.83	24

The reliabilities of the MCGJT and the DGJT were .84 and .83 Cronbach coefficient alpha (Tables 5 & 6). Since Nunnally (1972) proposes a minimal of .7 for Cronbach alpha value (as cited in Pallant, 2005), the reliabilities obtained are satisfactory with a slight superiority for the MCGJT.

Tables 7 and 8 reflect that the OGJT enjoys a comparatively higher reliability than the LGJT, and that overall, among the four response formats the OGJT holds the highest and the LGJT, the lowest reliabilities.

Table 7: *Reliability statistics for the OGJT*

Cronbach's Alpha	N of Items
.86	24

Table 8: *Reliability statistics for the LGJT*

Cronbach's Alpha	N of Items
.79	24

5.4 Validity Evidence (1)

With respect to the first research question, Alderson et al. (1995) postulate that internal correlations can provide pieces of evidence

for construct validity of a test. In this method, the lower are the correlations among the test components, the higher the construct validity. The logic lies in the fact that test subsets are expected to tap different aspects of the language ability the test aims at; therefore, their correlations are expected to be low to enhance the overall construct validity. This is also in line with Kline (1994) who views low correlations speaking to distinctness of traits. Tables 9 and 10 depict the internal correlations for the MCGJT and the DGJT.

Table 9: *Validity indices for the MCGJT*

	SU	**DO**	**IO**	**OPREP**	**GEN**	**OCOMP**
SU	1	.28**	.25**	.17	.19*	.17
DO	.28**	1	.62**	.54**	.41**	.34**
IO	.25**	.62**	1	.61**	.43**	.36**
OPREP	.17	.54**	.61**	1	.31**	.43**
GEN	.19*	.41**	.43**	.31**	1	.34**
OCOMP	.17	.34**	.36**	.43**	.34**	1

Note. SU = Subject; DO = Direct; IO = Indirect object; GEN = Genitive; OPREP = Object of preposition; OCOMP = Object of comparison.
**p < .01, two-tailed.
*p < .05, two-tailed.

Alderson et al. (1995), deem correlations of 0.3- 0.5 to be low enough; therefore, as can be noticed in Table 10, all correlations are satisfactorily low enough apart from six of them, i.e., DO and IO, DO and OPREP, IO and OPREP falling short of meeting the desirable low correlation.

Table 10: *Validity indices for the DGJT*

	SU	**DO**	**IO**	**OPREP**	**GEN**	**OCOMP**
SU	1	.21*	.22*	.25**	.43**	.09
DO	.21*	1	.56**	.35**	.22*	.44**
IO	.22*	.56**	1	.57**	.41**	.30**
OPREP	.25**	.35**	.57**	1	.41**	.16
GEN	.436*	.22*	.41**	.41**	1	.11
OCOMP	.09	.44**	.30**	.16	.11	1

Note. SU = Subject; DO = Direct; IO = Indirect object; GEN = Genitive; OPREP = Object of preposition; OCOMP = Object of comparison.
**p < .01, two-tailed.
*p < .05, two-tailed.

When compared with the MCGJT, the DGJT holds slightly lower correlations due to not having any components of 0.6 values (Table 10). This leads us to conclude that the DGJT yields mildly better validity indices.

Internal correlations, according to Alderson et al. (1995), can additionally be performed by the correlations between each subset

and the entire test reflecting the amount of contribution that specific component makes to the overall picture of construct validity. Thus, this type of correlation is expected to be higher, i.e., approximately about 0.7 or more. However, since in this regard the correlation between the subset and itself will also affect the results, it is customary to correlate the subset with the total test excluding the component in question.

Table 11: *Validity indices for both tests concerning the subsets and totals*

		SU	DO	IO	OPREP	GEN	OCOMP
MCGJT	Total	.46**	.76**	.81**	.76**	.64**	.66**
	Total minus self	.28**	.65**	.67**	.61**	.48**	.47**
DGJT	Total	.50**	.69**	.81**	.74**	.63**	.53**
	Total minus self	.34**	.55**	.66**	.54**	.47**	.31**

The correlations between the subsets, totals and totals minus self in both tests are more or less the same and the differences are quite negligible (Table 11). The correlations with total test fell within the range of 0.4-0.8 although they were expected to correlate around 0.7 as a satisfactory size of correlation with totals (Alderson et al., 1995).

5.5 Validity Evidence (2)

The validity, i.e., what the tests in differing response formats measured, was probed through a post-test questionnaire (See Appendix E).

Table 12: *Item frequency of the questionnaire for the OGJT*

Item	Frequency	Percent	Valid Percent	Cumulative Percent
a	8	18.6	18.6	18.6
b	13	30.2	30.2	48.8
c	6	14.0	14.0	62.8
d	10	23.3	23.3	86.0
e	5	11.6	11.6	97.7
f	1	2.3	2.3	100.0
Total	43	100.0	100.0	

In this questionnaire, according to Currie and Chiramanee (2010), item (a) reflects knowledge, (b) and (c) are both classified as partial knowledge, (d), which was added by the researchers, directly refers to the impact of response format, (e) indicates poor test-taking technique and item (f) is reflective of blind guessing. As can be observed in Table 12, the highest frequency belongs to (b) that is partial knowledge of over 50% and then to (d) which comes second suggesting that response format is the other influential factor.

Table 13: *Item frequency of the questionnaire for the LGJT*

	Item	Frequency	Percent	Valid Percent	Cumulative Percent
	a	10	23.3	26.3	26.3
	b	8	18.6	21.1	47.4
	c	7	16.3	18.4	65.8
	d	9	20.9	23.7	89.5
	e	3	7.0	7.9	97.4
	f	1	2.3	2.6	100.0
	Total	38	88.4	100.0	
Missing	System	5	11.6		
Total		43	100.0		

The items of the LGJT were the same as those of the OGJT apart from item (f) which refers to learning the answer after taking the first test. According to Table 13, unlike the OGJT, item (a), with a slight superiority to item (b), holds the highest frequency which is suggestive of knowledge. However, analogous to the OGJT, item (d) with a marginal difference is second to item (a). Overall, the results of the post-test questionnaire signify that the LGJT can better address the construct of the test; hence, it is rather superior to the OGJT. Besides, in both tests, item (d) has received a notable frequency holding the second place. It suggests that its influence as a systematic measurement error on validity of test scores is not negligible. It is also worthy of note that not finding their reasons among items listed, some participants wrote their justifications in the space provided. Relying on their intuition, being affected by stress and selecting mistakenly were among the reasons they mainly alluded to.

6. Discussion

With regard to the first research question, in accord with the study of Kazuo (2003), the MCGJT yielded better reliability indices and consequently higher determinacy since as Gass (1994) points out 'the issue of reliability cannot be separated from issues of indeterminacy' (p. 319). Indeterminacy, as she maintains, stems from learner's incomplete knowledge or lack of knowledge about that part of grammar in question. Sorace (1990) argues that there is incompatibility between dichotomous form of grammaticality judgment test and the learner's interlanguage nature (as cited in Kazuo, 2005). This, according to her, is reflected in the learner's random choice of 'correct' or 'incorrect'. Similarly, Gass and Selinker (2008) argue that in this vein when a learner marks a sentence as ungrammatical, one cannot be certain that his justifications conform with those of the researchers. To resolve this issue, they suggest requiring identification and provision of the correct form from the learner. This was the same technique incorporated in the DGJT in the current study.

However; the results still show superiority for the MCGJT, which can be attributed to the more chances test-takers are provided with in the MCGJT lessening the possibilities of random choices. In other words, in the MCGJT, not only do test-takers need to identify the statements as correct or incorrect, as they do in the DGJT, but also they have to choose one of the three choices presented under either of the headings and provide the correct form for the grammatically incorrect choice.

With respect to validity, the results of the present study are not in conformity with those of Kazuo (2003). The latter reported an improvement in the concurrent validity of the MCGJT, whereas the findings of this study revealed a small degree of superiority for the DGJT, which if considered negligible since they only differ in one pair of correlation, it can even be concluded that validity indices are equally satisfactory not being affected by the response format. The probable reason underlying this discrepancy can be attributed to the type of validity examined in these two pieces of research. Kazuo (2003) employed concurrent validity correlating the scores of the MCGJT and the DGJT with a C-test and TOEFL. Yet, this study attempted to investigate the construct validity due to the fact that, as Bachman (1990, p. 290) asserts, 'it subsumes content relevance and criterion relatedness' by empirically verifying hypotheses based on a theory of factors which influence test performance such as features of test method.

The reliabilities of the OGJT and the LGJT were found to be influenced by the response format since a notable difference in their Cronbach's coefficient alpha was observed. According to Henning (1987), fluctuations in the learner, scoring and test administrations as well as test characteristics are among the threats affecting reliability of test scores. Considering the tests in the current study, fluctuations in the learner, scoring and administrative factors are unlikely to exert major influences. This is owing to the fact that the factors mentioned were by and large maintained intact. Hence, test characteristics, as another influential factor, must have caused the existing variation. One of the well-recognized issues in this regard is the number of items. In the OGJT, each item was in fact comprised of three prompts which students needed to rank based on

their grammatical correctness. This, totally, enhanced the number of items test-takers were dealing with turning 24 items to 72 prompts, whereas in the LGJT they were required to answer to only 24 items.

Validity investigations revealed that response format is a potential threat to validity of test scores. This is well confirmed by the participants' choice of item (d), i.e., 'the format of the test confused me', of the post-test questionnaire for both tests as their second reason behind the inconsistency of their responses. The analysis of the questionnaire also disclosed slight superiority of the LGJT to the OGJT in that the former possessed higher frequency rates for tapping knowledge. Though marginal, this discrepancy could stem from being provided by choices they needed to rank since there is a strong possibility that some test-takers resort to some test-taking strategies rather than their knowledge to deal with the items. Several participants wrote in their questionnaires that they selected the correct option by excluding wrong ones or by comparing the options. However, to take the LGJT test-takers considered the sentences in isolation matching them with one of the points on the likert scale. Therefore, they could better rely on their knowledge rather than the test-taking strategies to reply.

7. Conclusion

The present research, in all, points to the shrouded influence response format leaves on reliability and validity of GJ tests. The findings encourage language testing professionals and researchers to reflect upon the problem of the high risk of chance-level errors in such tests, particularly in the most conventional formats, i.e., dichotomous and Likert. The researchers found that provision of more choices for each item can be effective in this regard resulting to higher reliability indices in the case of the MCGJT and the OGJT. Among the merits of the constructed MCGJT and the OGJT, their high reliability, satisfactory validity and ease of scoring can be mentioned; therefore, these test types have the potentials to be one of the major reliable tools in both language testing and SLA studies.

While transforming the original GJ test into tests with differing formats, the researchers came up with several 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. Third, the rather novel post-test questionnaire approach, which was adopted from Currie and Chiramanee (2010), merits the attention of interested researchers. The questionnaire directly and quickly elicited participants' reasons for their choice of responses in tests providing the researchers with clear insights about the observed inconsistencies.

The following insights can be of immense avail to future researchers. First, since NPAH arises out of implicational universals, i.e., they are data-driven, any exception can be considered as a drawback to their entity and consequently leading to their downfall (Cook & Newson, 1996). Taking this into account, it would be considerably worthy of research to examine whether the position of relative clauses on the accessibility hierarchy, as proposed by Keenan and Comrie (1977), is also observed by Persian learners. If conflicting, the results would also hint at those relative structures which are marked from the viewpoints of this

particular group of participants. Second, the findings of this study necessitate further investigation on the probable impact of response format on reliability and validity of other types of language tests. This would well eliminate uncertainties about this issue and would provide insights about the role that nature of language tests might play in this regard. Finally, 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).

References

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

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

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

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

Chapelle, C. (1988). Field independence: A source of language test variance? *Language Testing, 5*(1), 62-82.

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

Cook, V. J., & Newson, M. (1996). *Chomsky's universal grammar: An introduction.* (2nd ed). Cambridge, MA: Blackwell Publishing.

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.

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

Davies, W. D., & Kaplan, T. I. (1998). Native speaker vs. l2 learner grammaticality judgments. *Applied Linguistics, 19*, 183-203.

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

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.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). New York: Routledge/Taylor & Francis.

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

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

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

Kazuo, A. (2003). Development of multiple-choice grammaticality judgment tests. *JLTA Journal,* 6. Retrieved August 6, 2011, from http:// *homepage.mac.com/kintyre/research/articles/MC-GJ.pdf*

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

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

Kyzlinkova, L. (2007). *On communicative language competence, validity and different modes of administration.* Masaryk University, Czech Republic.

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

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

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

Pallant, J. (2005). *SPSS survival manual: A step-by-step guide to data analysis using SPSS version 12* (2nd ed.). Sydney: Allen & Unwin.

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

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

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

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

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

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

Yujie, J., & Wenxia, Z. (2007). Evaluating the construct validity of an EFL test for PhD candidates: A quantitative analysis of two versions. *Shiken: JALT Testing & Evaluation SIG Newsletter, 11*(1), 2-16

Appendices

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. woman……	The 2. I am taller than the woman.
3. Whom…….	3. The woman is taller than I know.

Appendix C. The Ordinal Grammaticality Judgment Test

1. I saw the man whom crossed the street.
1. I saw the man who crossed the street.
2. 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:

I chose this option in the second test because:
a) I was 100% sure that it was the best choice.
b) I was 50-99% sure that the choice I selected was correct.
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:
