Psychological Test and Assessment Modeling, Volume 59, 2017 (2), 207-228

The effect of gender and academic background on cloze and reading comprehension performance using G-Theory

Hamdollah Ravand[1] *& Sahar A Sardari*[2]

Abstract

This study employed generalizability theory to investigate the impact of academic background and gender on performance on cloze and reading comprehension performance. For this purpose, 5000 examinees who took the University Entrance Examination In Iran were studied. Results of the study showed that gender and academic background had negligible effects on reading and cloze test performance. However, the results showed a strong effect for the two-, three-, and four-way interactions of *person* with *item*, *gender*, and *academic background*. A notable finding in the present study was the highly similar effect of gender and academic background nd their interactions with item and person on both cloze and reading comprehension tests. The findings of the study may have implications for the construct equivalence of reading comprehension and cloze tests.

Keywords: generalizability theory, task type, cloze, reading comprehension, gender, academic background

[1] *Correspondence concerning this article should be addressed to:* Hamdollah Ravand, assistant professor, English Department, Faculty of Humanities, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran; email: ravand@vru.ac.ir

[2] Vali-e-Asr University of Rafsanjan

Introduction

Tests in educational testing are mostly used to provide information based on which decisions are made (Bachman, 1990). Therefore, it is of utmost importance that test results be a true representation of test takers' ability. In other words, the effect of factors other than test takers' ability such as gender, academic background and task type which are construct irrelevant factors, in Messick's (1989) words, should be kept at a minimum.

Many research studies have explored the effect of gender as a construct irrelevant factor on performance in reading comprehension (e.g., Bügel & Buunk, 1996; Gorjian & Javadifar, 2013; M.-L. Lee, 2012; Pae, 2004). However, the effect of gender on cloze test performance has been underexplored. There are some studies (e.g., Alderson & Urquhart, 1985; Birjandi, Alavi, & Salmani-Nodoushan, 2002; Chen & Graves, 1995; Clapham, 1998; Hale, 1988; Hung, 1990; Krekeler, 2006; Osman, 1984; Peretz & Shoham, 1990; Ridgway, 1997; Salmani-Nodoushan, 2003) that have explored the effect of background knowledge on reading comprehension performance. But the effect of background knowledge on cloze test performance has not been given as much attention (e.g., Al-Fallay, 1994; Chihara, Sakurai, & Oller, 1989; Sasaki, 2000; Sharafi & Barati, 2011; Tabatabaei & Shakerin, 2013).

There have been ongoing debates among researchers as to what a cloze test measures (e.g., Bachman, 1985; Chavez-Oller, Chihara, Weaver, & Oller, 1985; Chihara, Oller, Weaver, & Chavez-Oller, 1977; Jonz, 1990; McKenna & Layton, 1990). Bachman (1985) argued that before making any claim about construct validity of cloze tests, a distinction should be made between *fixed ratio* cloze tests where every n^{th} word of a text is deleted and *rational deletion* cloze tests in which words from a special category are omitted. Fixed ratio cloze tests have been shown to fail to assess test takers' reading comprehension ability (e.g., Bachman, 1985; Kintsch & Yarbrough, 1982) while rational cloze has been argued to be sensitive to test takers' reading comprehension ability (Alderson, 2005; Bachman, 1985; Greene, 2001; Kintsch & Yarbrough, 1982; Yamashita, 2003). Although, the documentation for the UEE does not provide any information regarding the rationale behind cloze deletions, close inspection of the test showed that most deleted words in the cloze test under investigation were interclausal/intersentential or depended on long range coherence pattern, which are typical deletions in a rational cloze test.

Paucity of and in cases mixed results of the effect of gender and academic background on cloze and reading comprehension test performance, calls for more studies. To this end, the present study aims to explore whether performance on reading comprehension and cloze tests is and to what extent affected by personal characteristics such as gender and academic background.

Significance of the study

The present study is significant from many aspects: (1) It can shed light on whether and to what extent personal characteristics affect cloze and reading comprehension test per-

formance hence provide evidence for Bachman's (1990) framework of test performance. (2) Similarity or differences in the pattern of the effect of personal characteristics may have implications for the equivalence (or lack thereof) the constructs measured by cloze and reading comprehension. Researchers are divided on whether cloze tests measure reading comprehension (Baghaei & Ravand, 2016). If we can show that both reading comprehension and cloze tests are affected equally by the same factors, it is likely that similar processes are involved in performing on the two tests. (3) The present study can also provide evidence as to the validity of Iranian university entrance exam (UEE) which screens applicants into English programs at Bachelor's level in a very tight competition. Finally, (4) the current study provides a step-by-step illustration of gneralizability theory (G-theory)application to language test data in order to investigate main and interaction effect of factors affecting test performance.

Literature review

In this section a review of the studies on the effect of gender and academic background on cloze and reading comprehension performance is provided. Then, application of the G-theory in studies on second and foreign language testing is reviewed.

Gender and academic background on reading comprehension and cloze test

A large body of research studies have been devoted to investigate sources of construct-irrelevant factors in reading comprehension and cloze tests. A number of studies have shown that test takers' gender can significantly affect their performance on a reading comprehension test. An early study conducted by Bügel and Buunk (1996) found significant differences across male and female groups. The differences were reported to be due to sex differences in reading habits. In another study, Pae (2004) found that female learners outperformed male learners in items classified as Mood/Impression/Tone, while male group performed better in Logical Inference items. M.-L. Lee (2012) also reported a significant difference in reading comprehension strategies used by male and female groups. According to Brantmeier (2003) there is a significant effect for interaction of readers' gender and passage content on second language reading comprehension performance.

Gender effects in other studies were found to be insignificant. For example, Abdorahimzadeh (2014) noted test takers' gender cannot affect their performance on a reading comprehension test. As to cloze test performance, Abdorahimzadeh found an absence of the relationship between topic interest and foreign language reading comprehension performance across gender groups. Phakiti (2003) also explored gender differences in strategy use in a foreign language reading comprehension test. He reported gender does not affect reading comprehension performance significantly but gender differences affect cognitive and metacognitive strategy use.

The literature on gender effect on second or foreign language cloze test performance is relatively scarce. Sharafi and Barati (2011) found no significant gender effect on test

takers' cloze test performance. Tabatabaei and Shakerin (2013) also reported no significant gender effect; however, test takers performed better on cloze tests with familiar contents.

The following points are notable about studies exploring the effect of gender on cloze and reading comprehension test performance: (1) The effect of gender on cloze test has been much less explored, compared to the effect of gender on reading comprehension test performance, (2) studies on the effect of gender on reading comprehension have come up with mixed results, and (3) few studies have compared the effect of gender on reading and cloze test performance.

Another potential cause of bias in evaluating test takers' performance is their background knowledge. Background knowledge that can be considered as topic or text familiarity, results from factors such as readers' experience or academic (educational) background. As to the effect of background knowledge on reading comprehension performance, nearly all studies done so far showed a facilitating (Alderson & Urquhart, 1985; Birjandi et al., 2002; Chen & Graves, 1995; Hale, 1988; Hung, 1990; Krekeler, 2006; Osman, 1984; Peretz & Shoham, 1990; Ridgway, 1997; Salmani-Nodoushan, 2003). However, Clapham (1998) did not see background knowledge as an advantage and argued for a threshhold level above which test takers can benefit from their language resources to compensate for a certain lack of background knowledge.

Studies especially devoted to academic background or test takers' field of study are relatively scarce and the findings are mixed. In some studies no effect was found for background knowledge. Carrell (1983) found that background knowledge does not affect nonnative readers performance significantly. Yet, other studies have shown test takers with different academic backgrounds perform differently. Hale (1988) found that test takers' academic background interacts with the text context; therefore, significantly affects their performance on a reading comprehension task of Test of English as a Foreign Language (TOEFL). Taillefer (2005) also reported a significant difference in reading comprehension performance and strategy use across groups with different academic literacy backgrounds.

A number of other studies have investigated the effect of background knowledge on test takers' cloze test performance. Sasaki (2000), for example, examined the effect of schemata activated by culturally familiar words on students' cloze test taking processes. The researcher found that those who were taking a cloze test with a culturally familiar text, understood the text better and tried to solve more items. In another study, Chihara et al. (1989) changed less culturally familiar words into more familiar ones in two cloze tests presented to Japanese participants. The results showed that participants performed significantly better on cloze tests with culturally familiar words. Al-Fallay (1994) also explored the effect of cultural familiarity on 74 Arab EFL learners' cloze test performance. He found test takers that completed culturally familiar cloze tests outperformed others presented with culturally unfamiliar texts. Sasaki (1993) also reported that test takers that performed better on the cloze tests used a greater amount of information to find the correct answer. According to Sasaki, high proficiency students put their already known information together and formed an appropriate schema that helped them perform better

on the cloze test. The results reported by Sasaki concur with Oller's (1995) hypothesis that appropriate schemata formation promotes cloze test performance.

The following points are notable about studies exploring the effect of academic backgrounds on cloze and reading comprehension test performance: (1) Although a set of studies have investigated the effect of academic background or field of study on test takers' reading comprehension performance, to the authors' best knowledge no study has investigated the effect of test takers' academic background on cloze test performance, (2) Studies on the effect of academic background on reading comprehension tests are scarce and have shown mixed results, and (3) No study, to the authors' best knowledge, has compared the effect of academic background on reading and cloze test performance.

Application of G-theory in language testing

G-theory is an extension of classical true score theory (CTS) and was first applied in language testing by Bolus, Hinofotis, and Bailey (1982). A notable advantage of G-theory lies in its capability to assess magnitude of interaction effects as well as that of the main effects. Through the estimates of interaction terms we can infer about the bias (or lack thereof) that each source of variance (referred to as "facet" in G-theory) may introduce into test scores. For example, if the magnitude of the interaction term between gender and task type is high, it shows that performance of test takers with different sexes differed on different forms of the test. It may be that, for example, females performed better on cloze whereas males outperformed females on multiple choice reading comprehension tests. Moreover, G-theory unlike CTS that deals with only one source of error at a time, accounts for different sources of error and reports the interaction between these sources as well (Gebril, 2013). According to Brennan (2001, P.3) "Generalizability theory liberalizes classical theory by employing ANOVA methods that allow an investigator to untangle multiple sources of error that contribute to the undifferentiated E in classical theory".

An observed score of, for example, person *P* for occasion *O* evaluated by rater *R* can be represented as:

$$X_{por} = \mu + \nu_p + \nu_o + \nu_r + \nu_{po} + \nu_{pr} + \nu_{or} + \nu_{por}$$

Where μ is the grand mean and ν denotes any one of the effects (main or interaction) or components.

The variance of the scores in Equation 1 over the population of persons, occasions, and raters is:

$$\sigma^2(\mathrm{X}_{por}) = \sigma^2(\mathrm{X}_p) + \sigma^2(\mathrm{X}_o) + \sigma^2(\mathrm{X}_r) + \sigma^2(\mathrm{X}_{po}) + \sigma^2(\mathrm{X}_{pr}) + \sigma^2(\mathrm{X}_{or}) + \sigma^2(\mathrm{X}_{por})$$

As Equation 2 shows, the total observed score variance is portioned into seven independent variance components.

The D study phase of G theory also makes it possible to evaluate the improvements that could be made over test reliability by making changes to conditions of facets of the study.

G-theory has been used to study issues related to second or foreign language testing. Brown (1999) used G-theory to study consistencies across items, subtests and languages in TOEFL tests. Solano-Flores and Li (2008) examined the dependability of academic achievement measures for English language learners. Karami (2012) also exploited G-theory to investigate the impact of persons, items, subtests, and academic background on dependability of scores from a high stakes language proficiency test. In another study, Karami (2013) applied G-theory to investigate the impact of students' gender on their performance in a high-stakes proficiency test. Lehmann (1983) explored consistency of writing compositions. Bachman, Lynch, and Mason (1995) used G-theory to investigate rater and task effect on the dependability of grammar ratings from a Spanish speaking test. Y.-W. Lee (2005) used G theory to study the relative effect of different speaking tasks and raters on the ESL students' oral performance and scores. Barkaoui (2007), as a part of his study, employed G-theory to investigate the effects of two different rating scales on EFL essay scores, rating processes, and raters' perceptions. Gebril (2009) explored comparability of scores in a writing test with two types of tasks: independent task and reading-to-write task. Huang and Foote (2010) explored the factors that can possibly affect holistic evaluation of English as a second language (ESL) graduate students' writings. Huang (2012) examined the accuracy and validity of the ESL writing scores. In a recent study, In'nami and Koizumi (2016) used G-theoy to investigate task and rater effect in L2 speaking and writing.

Mixed results and in some cases paucity of the studies on the effect of gender and academic backgrounds on reading comprehension, in general, and on cloze and multiple choice forms of reading comprehension, in particular, shows the need for further investigation of the effect of these factors as possible sources of variance in test takers' scores. Furthermore, G-Theory has been more applied to performance assessment especially to writing. Paucity of the application of G-theory in reading comprehension calls for more research. Moreover, since UEE is a high-stakes gate-keeping test which screens applicants into the English programs at Iranian state universities at Bachelor's level, investigation of its validity is warranted. To the best knowledge of the authors validity of the current test and the other high-stakes tests developed and administered by the Measurement Organization in Iran has been sporadically investigated (e.g., Barati & Ahmadi, 2010; Ravand & Firoozi, 2016).

The present study intends to answer the following research questions:

1) What are the relative contributions of gender and academic background to the total score variance and those obtained from the UEE cloze and multiple choice reading comprehension test scores?
2) What are the distributional characteristics and the reliability of the scores obtained from the total test and from each subtest representing the two methods (i.e., cloze and multiple choice reading) under investigation?
3) What would be the effect of changing the number of conditions of facets on score reliability?

Instrument and participants

The participants of the present study were a random subsample (N=5000; 66.8% females and 33.2% males) of examinees who took the University Entrance Examination for applicants into the Bachelor's English programs in Iran (UEE) in 2014. The test is prepared and administered by the Measurement Organization, which is responsible for the development and administration of almost all nation-wide university entrance examinations. Participants were from three different academic backgrounds: Humanity, Science, and Mathematics. Table 1 shows the distribution of participants' gender and academic backgrounds.

UEE is a high-stakes test which is used as the sole criteria to admit high school students into English programs at higher education institutes in Iran. It is composed of two sections: The first section which is composed of 25 items (12 grammar and vocabulary items, 5 cloze test items and 8 reading comprehension items) measures the same level of English knowledge as measured in the test for non-English fields of study and the second section is a measure of language proficiency which is composed of 70 items (10 grammar items, 15 vocabulary items, 5 sentence structure items, 10 language function items, 15 cloze test items and 15 reading comprehension items). All the items are multiple choice and the test is administered under a restricted standardized condition. For the second section of the test, the test takers have 105 minutes to answer the 70 items. For the purpose of the current study 15 cloze and 15 reading comprehension items of the second section of the UEE were explored.

Table 1:
The distribution of participants' gender and academic background

Academic background		Frequency	Percent
Mathematics	Female	1001	53.9
	Male	855	46.1
	Total	1856	100.0
Science	Female	1872	74.3
	Male	648	25.7
	Total	2520	100.0
Humanity	Female	465	74.5
	Male	159	25.5
	Total	624	100.0
Total	Female	3338	66.8
	Male	1662	33.2
	Total	5000	100.0

Data analysis

Regarding the G theory design, *person* (denoted as P) were defined as the *objects of measurement* that contribute to the *universe score* variance. *Item* (denoted as *I*), *gender* (denoted as *G*) and *academic background* (denoted as *B*) were considered as the facets of the study that contribute to error variance. In the study, persons were defined as random facets because levels taken into account were selected randomly from the population or universe of interest and each person in the universe had the same probability to be selected for the present study (Cardinet, Johnson, & Pini, 2011). Items were also considered as random facets. While the two other facets being investigated in the current study (gender and academic background) were fixed facets. In fixed facets all levels of the facets are taken into account and no sampling of levels occurs (Cardinet et al., 2011).

To address research questions, a generalizability study with I(P:G:B) design was conducted. Since persons belonged only to one of the categories of gender (either male or female) and only one of the categories of academic background (Humanity, Science, or Mathematics), persons were defined as nested within gender and academic background, in G-theories' notation P:G, and P:B, respectively. However, the two facets person and item were defined as crossed (P x I), since every test taker attempted every item. Figure 1 shows the variance partition diagram for the current study.

For the purpose of the analysis EDUG software program (Swiss Society for Research in Education Working Group, 2006), designed to conduct G-theory, was used. EDUG is a program which gives researchers estimates for the variance components, the relative and absolute error variance as well as generalizability and dependability coefficients. Most G-theory software programs are unable to deal with incomplete or unbalanced data (Cardinet et al. (2011). Therefore, to work with EDUG software, all missing responses were coded as 0, assuming that test takers didn't know the correct answer for the question. In addition, data were balanced in order to have equal sizes for all facets in the study.

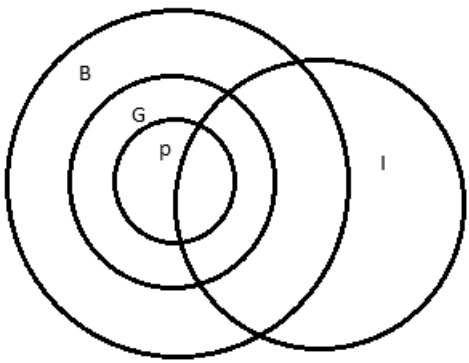

Figure 1:
The Variance Partition Diagram

Note. P: person, G: gender, B: academic background, I: item

To analyze the data, following Brennan (1983) and Brown (1999), thirteen generalizability studies were conducted to estimate the relative contribution of person, item, gender and academic background to the test scores: One total G-study considering reading comprehension and cloze tests together, two others examining reading comprehension and cloze tests separately, three separate G-studies taking into account the academic background of test takers in reading comprehension performance, two separate G-studies considering test takers' gender in reading comprehension task, and five separate studies exploring effect of different academic background and genders on cloze test performance. Brennan (1983) justifies such an analysis by stating: "when a population of objects of measurement is stratified with respect to several clearly defined subpopulations, it is almost advisable to conduct separate analyses for each subpopulation. In addition, an investigator may want to conduct a global analysis over subpopulations" (p. 93). The next step following a G study is a decision study (D study) or optimization phase. While G study's aim is to evaluate characteristics of a measurement, D study uses results obtained from a G study to make modifications to the procedure in order to improve it. In the D study phase changes to conditions of different facets will be done to evaluate improvements that could be made to the measurement design.

Results and discussions

To address the research questions of the study, this section is organized as follows: (1) The first section will present score distributions and classical reliability analyses of the total and balanced data for reading comprehension and cloze test sections, (2) The second section will present 13 separate G-studies that were conducted to investigate the relative contribution of person, item, gender and academic background and their interactions to score variance for the total test and each section of the test separately (reading comprehension and cloze test), and (3) The third section will present the subsequent D study to see what changes (i.e., increasing the number of facets levels, eliminating a facet, fixing a facet that was initially considered random or vice versa) can bring the best reliability estimates.

Score distributions and reliability analysis

In this section, distribution of scores and the classical reliability analysis for the total test and two subsets of test for both original data (n=5000) and balanced data (n=562) will be presented (Table 2).

As it is evident from Table 2, there is not much difference between descriptive statistics for the total data and the data balanced for the purpose of the study. Test takers' mean score on reading comprehension section was relatively higher than the mean score on the cloze section. The estimates of reliability, computed for both original data (n=5000) and balanced data (n=562) show the items making this scale hang together within each test and also in total quite well. Cronbach alpha values above .7 are considered to be ac-

Table 2:
Descriptive Statistics and Reliability Analysis

	Number of persons	Number of items	Mean	Std. Deviation	Cronbach's Alpha
Original data					
Reading comprehension	5000	15	13.1150	3.41131	.915
Cloze	5000	15	11.7910	3.03414	.915
Total test	5000	30	24.9060	4.51344	.946
Balanced data					
Reading comprehension	562	15	14.3086	2.57044	.889
Cloze	562	15	12.5617	3.13141	.912
Total test	562	30	26.8704	3.26492	.934

ceptable while values over .8 are preferable (Pallant, 2001). In other words, all items are measuring the same construct (Pallant, 2001). The spread of the test takers around their respective group means in both reading comprehension and cloze tests shows a very similar pattern.

Thirteen G-studies (items by persons nested within gender nested within academic background or I(PGB) design)

Tables 3, 4, 5, 6 summarize the percentage of variance explained by each facet and their interactions. In order to investigate the relative contribution of person, item, gender, and academic background, thirteen generalizability studies were conducted: one global analysis for all 30 items of reading comprehension and cloze tests, and twelve for each gender and academic background across the two task types (reading comprehension and cloze tests) separately. In the present study, the variance component contributing to the universe score is the person facet. The higher the variance component contributed by the person facet, the higher the tests reliability will be. Therefore, in the current study, large values for person facet and low values for other facets are preferable.

RQ1: *What are the relative contributions of persons, genders and academic backgrounds to the total score variance?*

Table 3 shows the result of the overall study with I(P:G:B) design. The largest variance component is associated with the residual (37.6% of the total variance). The residual contains the variability due to the interaction of person, item, gender, and academic background and other systematic or unsystematic error sources not included in the study. Triple interactions of PIB (17.6%), PGB (13.0%), and PIG (12.0%) yielded the next largest variances, respectively.

Other variance components accounted only for a small share of the total variance. The relatively small variance component due to the two-way interaction of PI (6.7%) indicated that, to some degree, persons' relative standing differed across items. Small values due to PB (4.7%) also indicated that persons from different academic backgrounds did not perform very differently. The percentage of the PG variance (3.2%) was relatively small, indicating that male and female test takers did not perform very differently. Among the main effects, items contributed to the score variance more than other variance components. The small variance due to items (3.1%) indicated that items are of about the same difficulty. Another variance component of interest is *person* facet which explains 1.5 percent of the total variance. This variance is considered as universe or true score variance. The relatively small variance component (1.5%) due to persons indicates that, contrary to what is expected of a norm referenced test, the UEE does not spread people out well. The variance explained by other variance components was zero. These findings taken together indicated that the test did not differentiate test takers well and although their performances were not affected by the main effect of the construct-irrelevant factors such as item and academic background, the interaction of these factors with each other and with the person facet greatly affected test results.

Table 3:
G-study results for the total test

Source of variation	VC estimate (% of total variance)
	The total test
P	1.5
I	3.1
G	0.0
B	0.0
PI	6.7
PG	3.2
PB	4.7
IG	0.2
IB	0.1
GB	0.1
PIG	12.0
PIB	17.6
PGB	13.0
IGB	0.2
PIGB	37.6

Note. VC = variance component; p= person; I=item; G=gender; B=academic background

Cloze and reading comprehension tests

To identify variance sources attributable to persons' abilities, item, gender, and academic background on cloze and reading comprehension test performance, two G studies were conducted. The results are presented in Table 4. Much aligned with the total tests, the four-way interaction of person, item, gender, and academic background yielded the largest variance component for both cloze (37.2%) and reading comprehension tests (37%).

For both reading comprehension and cloze tests, the next largest variance is related to the triple interaction of PIB (cloze: 18.0%, reading comprehension: 16.2%). There was also a large interaction effect for person, gender and academic background (PGB). It accounted for 12.6% and 15.1% of the shared variance in cloze and reading comprehension performance, respectively. The results provided evidence for a link between prior knowledge contributed to sex differences and test takers' foreign language reading comprehension performance. This finding is in line with Bügel and Buunk (1996) and Gorjian and Javadifar (2013). This difference could be due to differences in reading comprehension strategies used by male and female groups (M.-L. Lee (2012). According to Gorjian and Javadifar (2013), passages' topic can also provide information in terms of personal knowledge that may give one gender upper hand in comprehension, while making the comprehension difficult for the other gender. PIG accounted for the next highest

Table 4:
G-study Results for the Cloze and Reading Comprehension Tests

Source of variation	VC estimate (% of total variance)	
	Cloze	Reading comprehension
P	1.5	2.0
I	3.9	1.9
G	0.0	0.2
B	0.0	0.0
PI	6.5	6.4
PG	3.5	3.2
PB	4.3	6.4
IG	0.1	0.1
IB	0.0	0.1
GB	0.1	0.0
PIG	12.1	11.4
PIB	18.0	16.2
PGB	12.6	15.1
IGB	0.2	0.1
PIGB	37.2	37.0

Note. VC = variance component; p= person; I=item; G=gender; B=academic background

share of the variance (cloze: 12.1%, reading comprehension: 11.4%). Relatively small values for PI (cloze: 6.5%, reading comprehension: 6.4%) indicated persons' relative standing did not differ greatly across items. The two way interaction of PB also accounted for a small share of the variance in both reading comprehension and cloze tests (cloze: 4.3, reading comprehension: 6.4). The results indicated that test takers with different academic backgrounds performed almost similarly on both tests. Regarding the reading comprehension test, the findings are in line with Carrell (1983) while they are in contrast with some other studies (e.g., Hale, 1988; Taillefer, 2005). The cloze test results are in contrast with some other studies (e.g., Sasaki, 200; Chihara et al., 1989; Al-Fallay, 1994). The small share of variance by PG (cloze: 3.5, reading comprehension: 3.2) also shows that male and female test takers did not perform very differently. The results indicate that test takers' gender does not affect their performance on reading comprehension and cloze tests. Regarding the reading comprehension, this finding is consistent with some studies in the literature (e.g., Abdorahimzadeh, 2014; Phakiti, 2003). However, this finding runs counter to some other studies (e.g., Bügel & Buunk, 1996; Gorjian & Javadifar, 2013; M.-L. Lee, 2012; Pae, 2004). Cloze test results, in line with some other studies (e.g., Sharafi & Barati, 2011; Tabatabaei & Shakerin, 2013) indicated that gender does not affect test takers performance on a cloze test.

Among the main effects item (I) accounts for the greatest share of variance in the cloze test (3.9%) while the remaining small percentage of the variance was explained by person (1.5%). The contribution of other factors to the total variance was almost negligible. The small variance components due to person for both reading comprehension and cloze tests show that the test does not spread people out well. The relatively small variances for items shows that the items are almost of the same difficulty.

It is evident that the relative importance of each variance component to the score variance for both cloze and reading comprehension test almost mirrored those of the total test. Returning to Table 4, a comparison of the magnitude of different variance components across cloze and reading comprehension tests shows an almost similar pattern, which boosts the likelihood of the construct equivalence of the two tests.

Test taker's performance across different academic backgrounds

In order to examine the relative contribution of the person, item, and gender to test takers' performance across different academic backgrounds, 3 separate G studies for cloze and reading comprehension tests were conducted.

As it is evident in Table 5 the greatest variance component is the triple interaction of person, item and gender (PIG). The percentage of this variance for cloze test ranges from 43.5 for the Science group to 51.9 for the Humanity group. The percentage of this variance for the reading comprehension test also ranges from 44.7 for the Science group to 51.4 for the Humanity group. The two-way interaction of person and item (PI) yielded the second largest variance component for scores across different academic backgrounds for both cloze and reading comprehension tests, indicating that test takers' relative stand-

Table 5:
G-study Results for Performance across Groups with Different Academic Backgrounds

Source of variation	VC estimate (% of total variance)					
	Cloze			Reading comprehension		
	Math	Science	Humanity	Math	Science	Humanity
P	6.4	7.3	3.6	7.5	10.5	6.7
I	5.3	5.4	4.9	2.4	2.9	2.3
G	0.2	0.2	0.0	0.5	0.0	0.8
PI	23.5	22.5	28.1	22.5	21.6	24.7
PG	14.4	21.3	11.3	18.1	20.1	14.2
IG	0.1	0.7	0.2	0.1	0.6	0.0
PIG	50.2	43.5	51.9	48.4	44.7	51.4

Note. VC = variance component; p= person; I=item; G=gender; B=academic background

ing differed across items. Interaction of person and gender (PG) yielded the third largest variance component, suggesting that there was a difference in male and female test takers' performance.

Person (P), that represents true score variance, yielded the fourth largest variance component. For the cloze test, the true score variance accounted for 3.6 percent of the shared variance for the Science group, 6.4 percent for the Math group and 7.3 percent for the Humanity group. For the reading comprehension test, these values altered in a way that Humanity group, Math group and Science group explained 6.7, 7.5 and 10.5 % of the total variance, respectively. The effect of gender and its interaction with item was almost negligible in all groups. The findings indicated that items were of the same difficulty for male and female groups. The relative importance of each variance component to the test score variance across different academic backgrounds in both tests is almost the same.

A close inspection of the contents of Table 5 shows a very similar pattern of the effect of academic background for cloze and reading comprehension tests. The highest variance components of P and I occur in Science followed by Math and Humanity for both cloze and reading comprehension. PI has the highest variance in the Humanity group followed by Math and Science in both tests. However, PG has the highest variance for the Science followed by Math and Humanity. Finally, PIG had the biggest share of variance in Math followed by Humanity, and Science groups in both tests.

Test takers' performance across gender groups

As the next step, to examine the reliability of scores obtained from the two tests across different genders, two other separate G-studies for each test (cloze and reading comprehension tests) were conducted. As it is evident in Table 6, the largest variance component is the triple interaction between person, item, and academic background. For the cloze test, this variance ranges from 50.2 for male group to 57.8 for the female group. For the reading comprehension test, this variance ranges from 49.5 to 55.7 for the male and female groups, respectively. For the male group, the interaction of person and academic background (PB) yielded the second largest variance component (cloze: 20.6, reading comprehension: 24.6). Therefore, the results indicated that for the male group persons with different academic backgrounds performed more differently than the female group. However, for the female group, in both reading comprehension and cloze tests, the next highest observed variance is due to the interaction of person and item (PI). Therefore, females' relative standing differed across items more than female groups. The next highest variance components for the male and female groups were the interaction of person and item (PI) and the interaction of person and academic background (PB), respectively. Person (P), the object of measurement, yielded the next largest variance component to scores of all groups but cloze test scores of female group. Investigating groups with different genders, the relatively small variance component for person indicates that the tests did not spread test takers out.

Much aligned with other findings of the present study, the pattern of gender effects across both tests is very similar.

Table 6:
G-study Results for Performance across Groups with Different Genders

Source of variation	VC estimate (% of total variance)			
	Cloze		Reading comprehension	
	Male	Female	Male	Female
P	6.1	3.9	5.8	5.1
I	5.4	6.3	2.6	3.2
B	0.4	0.1	0.0	0.5
PI	17.4	19.7	17.2	19.4
PB	20.6	11.9	24.6	15.9
IB	0.0	0.4	0.3	0.1
PIB	50.2	57.8	49.5	55.7

Note. VC = variance component; p= person; I=item; G=gender; B=academic background

RQ 2: *What are the distributional characteristics and the reliability of the scores obtained from the total test and from each subtest representing two methods under investigation?*

The next question of interest was the dependability of the scores as measures of reading comprehension ability by investigating two sets of coefficients that are reported for each G-study: relative and absolute G coefficients. Exploring a norm-referenced test in the current study, the relative G coefficient is of importance. The relative G coefficient shows how much the relative standing of an individual is generalizable across different levels of the facets (Cardinet et al., 2011). As Table 7 shows, the relative G-coefficients estimates of almost all the cloze and reading comprehension analyses were higher than the minimum acceptable value (i.e., .8) reported by Cardinet et al. (2011). Therefore, it was shown that the two tests are reliable measures of test takers' reading comprehension ability. A notable pattern in Table 7 is that the relative G-coefficient of the two tests are very similar.

Table 8 shows the contribution of the relative error variance of the facets in studies. As it is evident in all three studies the interaction of person and item contributed 100.0 % to error variance. As Table 8 shows, gender and academic background do not contribute to the error variance since they are fixed facets. This information would be useful for a follow-up D-study analysis to show how to improve measurement precision.

Table 7:
Reliability Estimates for G studies

Source of variation	VC estimate (% of total variance)												
	Total test	Cloze					Reading comprehension						
		Cloze	Math	Science	Humanity	male	Female	RC	Math	Science	Humanity	Male	Female
Coef_G relative	0.87	0.78	0.80	0.84	0.76	0.84	0.79	0.81	0.83	0.88	0.80	0.83	0.80
Coef_G absolute	0.82	0.69	0.77	.80	0.62	0.80	0.69	0.79	0.82	0.87	0.79	0.81	0.77
Relative SE	0.025	0.038	0.067	0.060	0.068	0.049	0.058	0.032	0.052	0.055	0.053	0.047	0.043
Absolute SE	0.031	0.049	0.074	0.067	0.074	0.056	0.067	0.037	0.056	0.058	0.056	0.050	0.046

Note. VC= variance estimate; RC= reading comprehension

Table 8:
Contribution of relative error variance of facets for the studies that did not meet the acceptable reliability values

Source of variation	Relative error variance (% of total error variance)		
	Overall analysis of Cloze	Humanity group-cloze	Female group-cloze
P	….	….	….
I	….	….	….
G	….	….	….
B	….	….	….
PI	100.0	100.00	100.0
PG	0.0	0.0	….
PB	0.0	….	0.0
IG	…..	….	….
IB	…..	….	….
GB	…..	….	….
PIG	0.0	0.0	….
PIB	0.0	….	0.0
PGB	0.0	….	….
IGB	….	….	….
PIGB	….	….	….

Decision or optimization study (D-study)

RQ 3: *What would be the effect of changing the number of conditions of facets on score reliability?*

To answer the third research question, following the G-study a decision study (D-study) or optimization was conducted. Studies with G-coefficients below the acceptable value (.8) were taken and three D-studies were run. In all three studies (the overall analysis of the cloze test, cloze test performance of humanity and female groups) the interaction of person and item contributed 100.0 % to error variance. Therefore, in the optimization phase of the study the number of items was changed to see how the dependability could be increased. Referring to Table 9, one can see that the dependability of the scores will be increased if the number of items is increased to 30, 35 and, 25 in the analysis of the cloze test, for the Humanity group and female group, respectively. Therefore, more score reliability can be obtained by increasing the number of items.

A notable finding in the present study was that in all the G-studies the relative contribution of the facets and their interactions to the scores of reading comprehension and cloze tests showed a very similar pattern. As long as reading comprehension and cloze tests are

construct-equivalent there should be similar patterns of relationships between factors affecting test takers performance on these two different methods of reading comprehension. Therefore, findings of the current study, in line with some other studies (e.g., Bachman, 1985; Chavez-Oller et al., 1985; Chihara et al., 1977; Jonz, 1990; McKenna & Layton, 1990), provide support on the construct equivalence of cloze and reading comprehension tests. Thus, the present study found support for the proposal that rational deletion cloze tests are measures of test takers' reading comprehension ability.

A possible conclusion from the similar pattern of the effect of person, item, academic background, and gender on cloze and reading comprehension test performance is that the two forms measure similar traits. Trait equivalent studies of different test methods have either used the differences between the means obtained from the different methods (e.g., Shohamy, 1984) or the pattern of relationships of scores obtained based on each method and some external factors such as personal characteristics (e.g., Barati, Ravand, & Ghasemi, 2013). However, the mean differences may indicate only difficulty effects and, as Bennett (personal communication, June, 2013) argued "The strength and similarity of relations with external measures are much more critical indicators than similarity of the means." Before one can make a construct equivalence claim more personal characteristics should be studied. A possible source of the similarity of the effect of the studied factors on cloze and reading tests may be the equality of the test methods in both tests (i.e., both tests were multiple choice). Above all, neither patterns of the relationships with external factors nor comparison of the mean performance on cloze and reading comprehension is a conclusive evidence for trait equivalence of the two test forms. To ensure construct equivalence, *Construct representation* evidence (Embretson, 1983) should be sought. Construct representation research entails an examination of test responses from the point of view of the processes, strategies, and knowledge stores involved in the performance of test tasks.

Table 9:
D study

	Cloze				Cloze-Humanity				Cloze-female			
	G-study		optimization		G-study		optimization		G-study		optimization	
	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.	Lev.	Univ.
P	159	INF	159	INF	159	INF	159	INF	159	INF	159	INF
I	15	INF	30	INF	15	INF	35	INF	15	INF	25	INF
G	2	2	2	2	2	2	2	2	2	2	2	2
B	3	3	3	3	3	3	3	3	3	3	3	3
Coef_G relative	0.78		0.88		0.66		0.82		0.75		0.83	
Coef_G absolute	0.69		0.82		0.62		0.80		0.69		0.83	
Relative SE	0.038		0.033		0.068		0.042		0.058		0.050	
Absolute SE	0.049		0.042		0.074		0.045		0.067		0.042	

Results also provide evidence for the construct validity of the UEE. Since testing students' ability with the reading comprehension items or the cloze items made no difference in test takers' scores, the implication is that both test tasks measure the same construct.

The results also shed more light on Bachman (1990)'s framework. The findings of the present study indicated that gender and academic backgrounds as two categories of test takers' characteristic do not affect cloze and reading comprehension test scores significantly. However, the results showed that gender and academic background may not affect test takers' scores significantly, but their interaction effect with other factors can contribute to error variance.

The present study explored the effect of person, item, academic background, and gender on the cloze and reading comprehension sections of the UEE. To make claims about the validity of the tests more personal characteristics should be included into the study. To balance the data, in order to meet the requirement for working with EDUG, the number of the sample size under investigation was reduced to a great extent and that might limit generalizability of the results. Future studies may use either balanced designs or use software programs that work with imbalanced data.

References

Abdorahimzadeh, S. (2014). Gender differences and EFL reading comprehension: Revisiting topic interest and test performance. *System, 42*, 70-80.

Al-Fallay, I. (1994). Limiting bias in the assessment of English as a foreign language: the impact of background knowledge on the proficiency of Saudi Arabian students learning English as a foreign language. *Unpublished doctoral dissertation, University of New Mexico, Albuquerque*.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *Tesol Quarterly, 13*(2), 9-227.

Alderson, J. C. (1980). A process approach to reading at the University of Mexico. *ELT Documents: Special Issue. Projects in Materials Design*, 134-143.

Alderson, J. C. (2005). *Assessing reading*: Ernst Klett Sprachen.

Alderson, J. C., & Urquhart, A. H. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing, 2*(2), 192-204.

Bachman, L. F. (1985). Performance on cloze tests with fixed-ratio and rational deletions. *Tesol Quarterly, 19*(3), 535-556.

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

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language testing, 12*(2), 238-257.

Baghaei, P., Ravand, H. (2016). Modeling Local item Dependence in Cloze and Reading Comprehension Test Items Using Testlet Response Theory. *Psicologica. 37,* 85-104

Barati, H., & Ahmadi, A. R. (2010). Gender-based DIF across the subject area: A study of the Iranian National University Entrance Exam. *The Journal of Teaching Language Skills, 2*(3), 1-26.

Barati, H. Ravad, H. & V. Ghasemi (2013). Investigating Relationships among Test Taker's Characteristics and Response Formats in a Reading Comprehension Test: A Structural Equation Modeling Approach. *Iranian Journal of Language Testing 3*(2), 38-59.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86-107. doi: http://dx.doi.org/10.1016/j.asw.2007.07.001

Birjandi, P., Alavi, S., & Salmani-Nodoushan, M. (2002). Text familiarity, reading tasks, and ESP test performance: a study on Iranian LEP and non-LEP university students. *Unpublished PhD dissertation, University of Tehran. Available online at: www. geocities. com/nodoushan/HomePage. html (September 2005)*.

Bolus, R. E., Hinofotis, F. B., & Bailey, K. M. (1982). An introduction to generalizability theory in second language research. *Language Learning, 32*(2), 245-258.

Brantmeier, C. (2003). Does gender make a difference? Passage content and comprehension in second language reading. *Reading in a foreign language, 15*(1), 1.

Brennan, R.L. (1983). *Elements of generalizability theory.* Iowa City, IA: American College Testing Program.

Brindley, G. (1994). Task-Centred Assessment in Language Learning: The Promise and the Challenge.

Brown, J. D. (1999). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing, 16*(2), 217-238.

Bügel, K., & Buunk, B. P. (1996). Sex differences in foreign language text comprehension: The role of interests and prior knowledge. *The modern language journal, 80*(1), 15-31.

Cardinet, J., Johnson, S., & Pini, G. (2010). Applying generalizability theory using EduG. New York: Taylor and Francis.

Carrell, P. L. (1983). Three components of background knowledge in reading comprehension. *Language Learning, 33*(2), 183-203.

Chavez-Oller, M. A., Chihara, T., Weaver, K. A., & Oller, J. W. (1985). When are cloze items sensitive to constraints across sentences? *Language learning, 35*(2), 181-206.

Chen, H. C., & Graves, M. F. (1995). Effects of previewing and providing background knowledge on Taiwanese college students' comprehension of American short stories. *Tesol Quarterly, 29*(4), 663-686.

Chihara, T., Oller, J., Weaver, K., & Chavez-Oller, M. A. (1977). Are cloze items sensitive to constraints across sentences? *Language learning, 27*(1), 63-70.

Chihara, T., Sakurai, T., & Oller, J. W. (1989). Background and culture as factors in EFL reading comprehension. *Language testing, 6*(2), 143-149.

Clapham, C. (1998). The effect of language proficiency and background knowledge on EAP students' reading comprehension. *Validation in language assessment*, 141-168.

Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing, 26*(4), 507-531.

Gorjian, B., & Javadifar, M. (2013). Effects of Gender and Passage Content on Multiple-Choice Reading Comprehension Test. *Procedia-Social and Behavioral Sciences, 84*, 723-727.

Greene, B. (2001). Testing reading comprehension of theoretical discourse with cloze. *Journal of Research in Reading, 24*(1), 82-98.

Hale, G. A. (1988). Student major field and text content: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language testing, 5*(1), 49-61.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assessing Writing, 17*(3), 123-139. doi: http://dx.doi.org/10.1016/j.asw.2011.12.003

Huang, J., & Foote, C. J. (2010). Grading Between the Lines: What Really Impacts Professors' Holistic Evaluation of ESL Graduate Student Writing? *Language Assessment Quarterly, 7*(3), 219-233.

Hung, C. (1990). The effects of pre-reading instruction on the comprehension of text by ESL readers. *The English Teacher, 19*.

In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language testing, 33*(3), 341-366. doi: doi:10.1177/0265532215587390

Jonz, J. (1990). Another turn in the conversation: What does cloze measure? *Tesol Quarterly, 24*(1), 61-83.

Karami, H. (2012). The relative impact of persons, items, subtests, and academic background on performance on a language proficiency test. *Psychological Test and Assessment Modeling, 54*(3), 211.

Karami, H. (2013). An investigation of the gender differential performance on a high-stakes language proficiency test in Iran. *Asia Pacific Education Review, 14*(3), 435-444.

Kibby, M. W. (1980). Intersentential processes in reading comprehension. *Journal of Literacy Research, 12*(4), 299-312.

Kintsch, W., & Yarbrough, J. C. (1982). Role of rhetorical structure in text comprehension. *Journal of educational psychology, 74*(6), 828.

Krekeler, C. (2006). Language for special academic purposes (LSAP) testing: the effect of background knowledge revisited. *Language testing, 23*(1), 99-130.

Lee, M.-L. (2012). A study of the selection of reading strategies among genders by EFL college students. *Procedia-Social and Behavioral Sciences, 64*, 310-319.

Lee, Y.-W. (2005). Dependability of scores for a new ESL speaking test: Evaluating prototype tasks. *Monograph Series MS-28). Princeton, NJ: Educational Testing Service*.

Lehmann, R. (1983). Rating the quality of student writing: findings from the IEA study of achievement in written composition. *Language testing: new openings. Jyväskyla: University of Jyväskyla*, 186-204.

Markham, P. L. (1985). The rational deletion cloze and global comprehension in German. *Language learning, 35*(3), 423-430.

McKenna, M. C., & Layton, K. (1990). Concurrent validity of cloze as a measure of intersentential comprehension. *Journal of educational psychology, 82*(2), 372.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Osman, S. (1984). Effects of prior knowledge on ESL reading. *Reading in Asia: The first yearbook of CCA*, 43-61.

Pae, T.-I. (2004). Gender effect on reading comprehension with Korean EFL learners. *System, 32*(2), 265-281.

Pallant, J. (2001). *SPSS survival manual: A step-by-step guide to data analysis using SPSS for Windows (Version 10)*: Allen & Unwin.

Peretz, A. S., & Shoham, M. (1990). Testing reading comprehension in LSP: Does topic familiarity affect assessed difficulty and actual performance. *Reading in a foreign language, 7*(1), 447-455.

Phakiti, A. (2003). A closer look at gender and strategy use in L2 reading. *Language learning, 53*(4), 649-702.

Ravand, H., & Firoozi, T. (2016). Examining construct validity of the master's UEE using the Rasch model and the six aspects of the Messick's framework. *International Journal of Language Testin, 6* (1).

Ridgway, T. (1997). Thresholds of the background knowledge effect in foreign language reading. *Reading in a foreign language, 11*, 151-166.

Salmani-Nodoushan, M. A. (2003). Text familiarity, reading tasks, and ESP test performance: A study on Iranian LEP and non-LEP university students. *The reading matrix, 3*(1), 1-14.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language testing, 17*(1), 85-114.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, 229-255.

Sharafi, F., & Barati, H. (2011). The effect of Iranian EFL learners' cultural knowledge on their performance on cloze tests.

Solano-Flores, G., & Li, M. (2008). Examining the Dependability of Academic Achievement Measures for English Language Learners. *Assessment for Effective Intervention, 33*(3), 135-144.

Swiss Society for Research in Education Working Group. (2006). *EDUG user guide.* Neuchatel, Switzerland: IRDP.

Tabatabaei, O., & Shakerin, S. (2013). The effect of content familiarity and gender on EFL learners' performance on MC cloze test and C-test. *International Journal of English Language Education, 1*(3), 151-171.

Taillefer, G. F. (2005). Foreign Language Reading and Study Abroad: Cross-Cultural and Cross-Linguistic Questions. *The Modern Language Journal, 89*(4), 503-528.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing, 20*(3), 267-293.