Journal of Modern Research in English Language Studies
Vol. 6, No. 1, 2019, 1-29

An Investigation into Item Types and Text Types of Reading Comprehension Section of Iranian Ph.D. Entrance Exams Using G-theory

Masoumeh Ahmadi Shirazi[1], Seyyed Mohmmad Alavi[2], Hossein Salarian[3*]

[1] Associate Professor, Faculty of Foreign Languages and Literatures, University of Tehran, Iran, *ahmadim@ut.ac.ir*
[2]Professor, Faculty of Foreign Languages and Literatures, University of Tehran, Iran, *smalavi@ut.ac.ir*
[3*] Ph.D. Candidate, Faculty of Foreign Languages and Literatures, University of Tehran, Alborz Campus, Iran, *hossein_salarian@yahoo.com*

Abstract

This study investigates some problems of Ph.D. applicants in their entrance exams in the case of answering the reading comprehension questions. To this end, the researchers considered the *item types* and *text types* as main effects and *their interaction* effect using generalizability theory for examining the variability; the answer sheets of a mock-test from 321 applicants, from all parts of Iran, enrolled in an institute were randomly selected. Using a partially nested design of G-theory in the GENOVA program, the researchers identified five variance components in the two different passages with distinct items and investigated various sources of error that are involved in the measurement process. The results of the study showed that the main effect for items cannot be separated from the interaction between items and texts, and clarified that the *items* facet had a noticeable amount of variance, and therefore, they impacted the applicants' performance. However, the results of D-studies showed that the main effect for text types was zero, and both texts were at the same level of difficulty. Also, the persons had effects on the texts in their interaction. This study can motivate the researchers, test developers, and test designers to consider their work more carefully.

Keywords: D-study, Generalizability Theory, G-study, Ph.D. Entrance Exam, Reading Comprehension

Received 4 March 2019 | Accepted 03 July 2019
Available online 01 August 2019 | DOI: 10.30479/jmrels.2019.10591.1326

1. Introduction

In theory, answering the reading comprehension questions is viewed as a classification mechanism; this sort of mechanism determines the type of question and the related sources of information (Cerdan, Vidal-Arbarca, Martinez, & Gil, 2009; Rouet, Vidal-Abarca, Erboul, & Millogo, 2001). Indeed, determining the related sources of information entails empowering nodes in a knowledge network which is composed of not only textual information but also background knowledge (van Steensel, Oostdam, Amos van & Gelderen, 2013). Along this line, Rouet et al. (2001) contended that the improvement of this activation process depends on the nature of a test item which is related to the number of nodes used and their availability.

Moreover, Rupp, Tracy, and Choi (2006) argued that answering the passages with multiple-choice (MC) items is definitely different from the time of answering them in non-testing settings in which the readers do not think about MC items. In addition, Kendeou, McMaster, and Christ (2016) considered reading comprehension as complicated and with several components in a unit of language. Rupp, et al. (2006) also regarded 'comprehension' as a complex unit of language in which the design of items and the selection of passages for the assessment of reading comprehension definitely emphasize this construct. To put it differently, according to Pearson, Valencia and Wixson (2014), we have involved in difficulties and problems with reading comprehension assessment separately and wholly since thirty years ago.

On the other hand, Mostow, Huang, Jang, Weinstein, Valer, and Gates (2016) believed that MC tests psychometrically give a better estimate of reliability and have the advantage of ease of scoring. Moreover, if we observe Iranian Ph.D. applicants' general English performance on the entrance exams, we will notice that their performance on the reading comprehension section is weak. Therefore, one problem to be considered with these exams is the type of test items and/or their levels of difficulties. One research on the performance of testees based on the type and level of test items was conducted by Vidal-Abarca, Gilabert, and Rouet (1998); they concluded that the performance of testees on ‘high-level questions’ is very low; in these types of questions, the testees are required to focus on broader senses of concepts and inferential questions in contrast to 'low-level questions' which focus on a single concept and search for the answers of the questions in smaller passages.

The other problem which can be considered is the text types or genres, especially scientific and philosophical ones that Schoonen (2005) considered the "possible sources of variation" (p. 2). To be specific, some scholars argued (e.g., Brown 2011; In’nami & Koizumi, 2015), another

problematic factor in reading comprehension impacting the chances of success of test-takers is the interaction of the aforementioned problems in which the text types, items and even sometimes bias cause problems for these applicants or in some other cases in Linn's (1981) term they pollute the test scores with construct-irrelevant test score variance.

Generalizability theory (G-theory; Brennan, 2001; Cronbach et al., 1972) is one of the approaches supplying a framework for identifying the flawed items or text types, therefore, estimates their accuracy of precision (Brennan, 2001). However, Fan and Sun (2014) argued that many scholars do not understand the G-theory and its strength over traditional forms of reliability estimates (e.g., test-retest reliability, Cronbach's coefficient α). They also added that all other seemingly diverse reliability coefficients are involved in this theory.

Regarding the aforementioned problems, this study attempts to answer the following research questions in relation to the use of text types and their items in reading comprehension section of general English on Ph.D. Entrance Exams in Iran:

1. Does the item type in reading comprehension affect the performance of the applicants on Ph.D. entrance exams?

2. Does the text type affect the performances of the test-takers or applicants on Ph.D. entrance exams?

3. Does the interaction of genres and item types have any impact on score reliability?

2. Literature Review

2.1. Generalizability Theory

From historical point of view, scholars in language testing have used different procedures for evaluation (e.g., ANOVA, factor analysis) to obtain data for testing (Bolus et al., 1982). Basically, the theoretical construction for calculations was classical test theory (Brennan, 2001a, 2001b) particularly in discovering the variability between scorers in performance assessment (Huang, 2008).

The simplest measurement theory is classical test theory (CTT), and has been to a large degree applied in order to identify reliability of measurements (Bachman, 2004; Eason, 1991). In this theory, the observed score, which is obtained by any one individual, consists of a true score (T) and a random error (E). A true sore shows the real performance of an examinee and is totally reliable (Kieffer, 1998). Also, it relies on the individuals involved in it but not on the conditions of observation (Kane,

2010). However, an observed score which may not be adequately reliable is considered for the performance (Kieffer, 1998); it relies on not only the person but also the specific observation. Also, according to Brennan (2011), error scores are random components. Because the error changes and differs from diverse observations and different persons, the observed scores also differ in these conditions (Kane, 2010). Further, this theory is often used for unlimited participants in which each of them can be observed in many times (Cronbach, 2004).

Moreover, Bachman (2004) maintained that in CTT all errors of measurement are unpredictable and unsystematic, and the test takers' true ability is shown in their scores in a reliable test, but these scores don't display the measurement errors. In addition, Kane (2010) considered the true scores in classical test theory as variables which are identified, or produced in order to show the fixed component of the observed score for each person; however, this variable or construct has not fixed values in different observations.

On the other hand, Cronbach and his colleagues developed a general model for estimating the relative impacts of diverse sources of variation in test scores (Cronbach et al., 1963; Cronbach et al., 1972; Gleser et al., 1965); their developed model is known as generalizability theory (G-theory), which used the framework of factorial design and the analysis of variance (ANOVA). This theory came into language testing in 1982 (Bolus et al., 1982). Then, this theory developed little by little and made inroads into various domains of language testing. It is an approach to estimate measurement precision for situations in which measurements have multiple sources of error (Cardinet et al., 2011).

The current investigations on G-theory (e.g., Brennan 2010; Cardinet et al, 2009; Shavelson & Webb, 1991) have made this model more familiar to researchers and practitioners. A score in this model, according to Bachman (2010), is a sample taken from a hypothetical universe of possible measures. Therefore, scores are treated as dependable in G-theory when we get accurate inferences about the universe of permissible observations (Shavelson & Webb, 1991). Bachman (2010) maintained that we can generalize individuals' performances to other contexts based on their performances on a test in the way that there is a direct association between the test score as a sample of performance and generalizability; this indicates that we can consider reliability as a matter of generalizability which defines the universe of measures based on a given score. In Messick's (1989) term, generalizability is 'a component of construct validity' that could be explained via reliability or transfer (p. 250). In his view, the generalizability concept can be considered in terms of either stability of scores or transfer of test tasks to a larger domain – Bachman and Palmer (1996) call this domain 'target language use'.

All in all, G-theory has been applied in several studies, among them are the influence of tasks and scorers on L2 speaking (Lee, 2005; In'nami & Koizumi, 2015; Stansfield et al., 1992b), the impact of tasks on L2 writing research (Barkaoui, 2007; Wang, 2010), the effect on the person-by-task interaction (Huang 2009; In'nami & Koizumi, 2015), the impact of genre on the generalizability of writing scores (Bouwer et al., 2015; Schoonen, 2005; Van den Bergh, et al. 2012), the study of the reliability and validity of EFL/ESL writing marks (Gebril, 2010; Huang, 2008, 2012; Han & Ege, 2013; Huang & Foote, 2010; Huang & Han, 2013; Swartz et al., 1999), the investigation into the accuracy and validity of the writing scores designed for ESL learners (Huang, 2012), the influence of scoring methods on the reliability and variability of EFL writings (Huang & Han 2013), the investigation into the dependability and validity of a criterion-referenced test (Kunnan,1992), and identification of the suitable model used in the applications of G- theory (Zhang & Lin, 2016).

Several scholars discussed various advantages of G-theory (e.g., Bachman, 1990; Brennan, 2001; Fan & Hansmann, 2015; Fan & Sun, 2014; Shavelson & Webb, 1991; Swartz et al., 1999; Thompson, 2003; Webb & Shavelson, 2005; Vispoel et al., 2018). In spite of the vast advantages of G-theory, some scholars contended that it has some shortcomings and limitations (e.g., Shavelson & Webb, 1991; Strube, 2002; Webb et al., 1988); these defects are:

1. Due to its technical development, many researchers cannot easily realize it.

2. It may result in coefficients which are related to the specific sample for doing the research and, therefore, generalizing the results of the study to another population is restricted.

3. We need significant attempt in its design, data collection, and analysis and estimation of error sources.

On the other hand, Brennan (2001) argued that G-theory is the offspring of Classical Test Theory (CTT) and Analysis of Variance (ANOVA) and, therefore, considered them as the "parents" of G-theory. This mixture or marriage copes with the CTT inability to divide sources of variance. He further added that the use of ANOVA in G-theory "liberalizes" (p.3) CTT so that the researcher can investigate easily multiple sources of error involved in undifferentiated errors of CTT. Moreover, "generalizability investigations are helpful both for comprehending the relative significance of different sources of error and for formulating the organized measurement procedures" (Brennan, 2001, p. 4). The summary of important contributions in the history of G-theory is shown in Table 1.

Table1

A Summary of Significant Contributions in the History of G-Theory

Year	Researcher	Contribution
1955, 1959	Lord	Used the findings of Ebel's (1951) paper about rater's main impacts in his papers on conditional standard errors of measurement and reliability under the assumptions of the binomial error model. His studies were finally considered the distinction between relative and absolute errors in G-theory.
1960 to1965	Cronbach, Gleser and Rajaratnam	Had developed univariate G-theory.
1972	Cronbach, Gleser, Nada, Rajaratnam	Published the book *The Dependability of Behavioral Measurements: Theory of Generalizabilty for Scores and Profiles.*
1976 to1981	Cardinet, Tourneur, and Allal	Emphasized the *principle of symmetry* of the G theory; i.e., the function of components other than testees as the objects of measurement.
1983	Crick and Brennan	Developed *GENOVA,* a computer program, for the analysis of G-theory.
1989	Feldt and Brennan	Worked on reliability and G-theory.
1991	Shavelson and Webb	Published the book *Generalizabilty Theory: A primer.*
1992	Brennan	Worked on GT for classroom use in the *Educational Measurement: Issues and Practice*.
1993	Ferrara	Argued that G-theory has a crucial function in all aspects of educational assessment.
1998	Lynch and McNamara	G-theory makes it possible for researchers to identify the facets related to the assessment context of interest(i.e., the universe of acceptable observations).
2000	Marcoulides	The facets which can vary without making the observation unacceptable or unreliable are very important.
2005	Sudweeksa, et al.	In G-theory, the meanings of three terms should be clarified. They include: (a) facet, (b) interaction, and (c) reliability.
2013	Han and Ege	There is a direction towards the application of G-theory in performance assessment.
2014	Fan and Sun	G-theory depends on ANOVA for dividing the total score variance.
2014	Lin and Zhang	Studied the reliability of the judgments of reviewers about the indicators of language performance in terms of the number of reviewers.
2015	Wu and Tzou	Considered the benefits and effectiveness of multivariate G-theory in identifying the accuracy of cut scores in practical applications of standard setting procedures.

Table 1 (Continued)

2016	Vispoel, Morris, and Kilinc	In a G-theory ANOVA design, *persons* show the object of measurement and sources of measurement error show the facets of interest. In turn, these facets lead to the terms *universe score* and *G- coefficient*.
2018	Vispoel et al.	Clarified different manners in which G-theory can overcome defects of conventional reliability coefficients.

2.2. Generalizability and Decision Studies

According to Brennan (2001), G-theory differentiates a generalizability (G-) study from a decision (D-) study which Shavelson and Webb (1991) considered as two stages in the application of G-theory and Vispoel et al. (2018) regard them a two-tiered process in the analysis of G-theory. Moreover, Brennan (2001) contended that when a researcher wants to divide error into component parts and calculate the important sampling configurations, s/he does this process through these two kinds of studies.

Indeed, according to Smith and Kulikowich (2004), a G-study includes the universe of acceptable observations (e.g., any item, genre, or occasion), and is used to obtain calculations of variance components for the universe of admissible observations. These variance components are assumed to be generalizable. This information makes it possible for us to make various modifications of the initial G-study design in the next stage which is called a D-study. They further added that in D-study we apply the calculated variance components from the G-study to identify components of variance for other designs of researches (different number of items, genres, etc. than in the original G-study) that reflect the universe of generalization. In the same line, Brennan (2001) contended that although G-theory emphasizes the interpretation of variance components and measurement error, it supplies summary coefficients and thus he considered them as reliability coefficients rather than reliability-like coefficients. However, Everitt and Howell (2005) argued that this measurement theory differentiates between two reliability-like summary coefficients; i.e., generalizability coefficient (G- coefficient) and phi coefficient (symbolized as F). Still, some scholars such as Sudweeksa et al. (2005) argued that G-theory supplies four summary statistics which are absolute error variance, relative error variance, the *g*-coefficient for relative decisions, which Vispoel et al. (2017) considered similar and sometimes identical to reliability coefficient of CTT for norm-referenced uses of scores. The last is the phi-coefficient for absolute decisions, which was referred to as ‘an index of dependability’ by Brennan and Kane (1977).

2.3. Text Types and Item Types

According to Vidal-Abarca, et al. (1998), the most common types of text types that are used reading comprehension are: narrative, expository, argumentative, and descriptive. Within these kinds of text, we can expose different items in reading comprehension test, from explicit to implicit and inferential items. Marshall (1998) believed narrative text is a type of writing in which the objective is to tell an event or a story; this type of writing, which is effective for elaborating details and sequencing them, usually applies the mode of descriptive writing. Some instances of this kind of writing are novel, short story, biography and anecdotes. According to Morrell (2006), re-creation, invention, or visually presentation a person, place, event, or action is the objective of description is whereby the reader can imagine that which is being described. Instances of descriptive are journal writing, and poetry. Also, Morrell (2006) believed that the goal of expositive text is to explain, inform, or even describe information by suggesting and giving an opinion, and suitable discussion. It also supplies background information for instruction or amusement. And, according to Selgin (2007), argumentative type of text, which is also known as persuasive, examines a subject by collecting and evaluating the data so as to show the validity of an idea, or perspective via appropriate reasoning, and discussion.

3. Method

3.1. Participants and Context

The participants of this study were 321 Ph.D. applicants in Iran enrolled at Modarresan-e- Sharif institute from all over the country. They were part of a larger population with different majors. Their ages ranged from 24 to 40. All the participants spoke Farsi as their formal language and logically are considered on average at intermediate level of proficiency because of passing some courses (general and especial ones) in B.A. or B.S. and M.A. or M.S. in English, besides passing some courses in high school. It should be mentioned that those applicants whose majors were teaching English language, English literature, English translation, and linguistics were excluded from this study due to the fact that the purpose of this study was to consider the performances of those applicants who took only general English. Since the sample size is far smaller than the universe size, the researchers obtained the sample randomly. No information regarding their age, names, average score, and the socioeconomic status was provided by this organization.

3.2. Design and Instrument

The effectiveness of data obtained from a G-study is essentially associated with the design of the study (Brennan, 2001). According to

Shavelson and Webb (1991), since we can replace the sample conditions for any other sets of conditions from the universe which are of the same size, the facet can be categorized as random. Also, all the items and text types were the same for all these applicants and they could answer all of them. Therefore, all the conditions of one facet (e.g., items) are in a *crossed* design with all conditions of another source of variation (e.g., persons) in this measurement. On the other hand, each of the two passages had five fixed items not involved in the other one. Therefore, the design of the study is not in the complete form of crossed one, but is in the form of the partially nested one. The object of measurement is persons (*p*), or the applicants, and the facets are items *(i)* and text types *(t)*, and the item facet is nested within text types. Thus, the researchers used a $p \times (i: t)$ design.

All in all, one mock-test of 'General English' similar in form and format of Ph.D. Entrance Exams in Iran was presented to these applicants by Modarresan-e- Sharif Institute. Basically, the 'General English' test in Ph.D. Entrance Exams in Iran consists of 30 items including three sections of grammar, vocabulary and reading comprehension (RC). Moreover, the text types for these exams are mostly either expositive or argumentative. The answers in answer sheets were analyzed with the GENOVA program.

3.3. Procedures and Data Analysis

This study was conducted in two phases. At first, the researchers determined whether the items are functioning appropriately using item facility and item discrimination. Then, the G-study and D-study were done using GENOVA software (Crick & Brennan, 1983). In this way, variance components for each of the sources of variability were determined in order to estimate variance components (the different sources of variance) of the main facets, the object of measurement and their interaction, and the relative sizes of the variance components (VCs), then they calculate the reliability of observed scores in the contexts needing relative decisions, and absolute decisions.

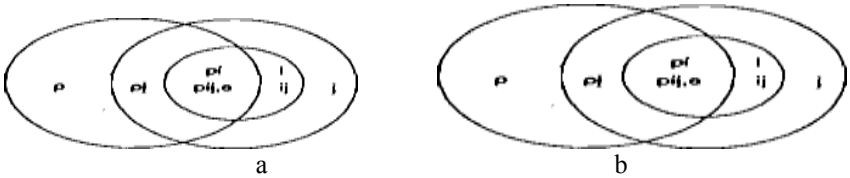
Certainly, the variance component for items $(\sigma^2 i)$ is confounded with the variance component for the item-by-text interaction ($\sigma^2 it$). Therefore, the variance component for these confounded effects is shown as $\sigma^2_{i,it}$. In this context, what the researchers wanted to explore was the true differences among the persons (P). The other components are known as "facets," which are potential sources of measurement error (Fan & Sun, 2014). Admittedly, the combined variance component is $\sigma^2 pi,\ pit,e$. As a result, the $p \times (i: t)$ design, has five variance components that can be investigated independently (see Table 2).

Table 2

Sources of Variability in the Two-facet RC Section Within Partially Nested Design

Source of Variability V	Type of Variation	Variance Notation
Persons (p)	Universe-score variance (object of measurement)	σ^2p
text types (t)	Constant effect for all persons due to inconsistency of genres	σ^2t
Items: text types (i:t)	Each item is nested within each given genre	σ^2i,it
p × t	Inconsistencies from one genre to another in persons' performances	σ^2pt
p i :t, e	Residual consisting of the three-way interaction and remaining unmeasured sources of error	σ^2pi, pit.e

The variance component for *persons* is the universe-score variance; it demonstrates the amount of systematic variability. Because items are nested within texts, it is impossible to separate the item main effect from the interaction between items and texts. As Shavelson and Webb (1991) argued, all variance components in nested facets cannot be calculated independently; the items in one text type differ from the items in another text type. Therefore, the researchers interpret the variance component for those combined effects of the total variance. The residual part of the total variance shows the amount of variation produced by these confounded sources of variation. Figure 1 shows the Venn diagram for this two-facet partially nested design with five sources of variance and the corresponding variance components.



a b

(a) Sources of variability (b) Variance components

Figure 1. Venn Diagrams for a Two-Facet, Partially Nested p × (i: t) Design

According to Shavelson and Webb (1991), a variance component (σ^2) determines the amount and degree of importance each source has in the measurement. A person's universe score, μ_p, is defined as the *expected value* (E) of the random variable, *Xpit,* across items and text types. The observed score in this type of measurement can be divided as follows:

$$
\begin{aligned}
Xpit = \ & \mu && \text{[grand mean]}\\
& + (\mu p - \mu) && \text{[person effect]}\\
& + (\mu t - \mu) && \text{[text type effect]}\\
& + (\mu it - \mu t) && \text{[item nested in text effect]}\\
& + (\mu pt - \mu t - \mu p + \mu) && \text{[person-by-item effect]}\\
& + (X_{pit} - \mu_{pt} - \mu it + \mu t) && \text{[residual effect]}
\end{aligned}
$$

Again, the grand mean, a constant for all persons, puts the score on the specific scale of measurement. Since it is a constant, its variance is zero. The observed-score variables vary from each other based on their levels, apart from for the grand mean, μ. The total variance over the universe and population is the sum of the variance components for the five effects:

$$\sigma^2(X_{pit}) = \sigma^2 p + \sigma^2 t + \sigma^2 \text{i,it} + \sigma^2 pt + \sigma^2 \text{pi,}pit.e$$

The residual has also a mean of zero and a variance shown as σ^2 pi,*pit,e*. The *p* × *t* effect shows that not all persons find the same texts easy or difficult. The ***e*** effect shows, to some extent, unsystematic or random error sources.

By referring to Sudweeks' et al. (2005) view, the researchers compared the relative size of the estimated variance components so as to identify the troublesome sources of variation, and determine the unwanted inconsistencies in obtaining good marks in the section of reading comprehension. Fan and Sun (2014) argued that variance components are *estimated*; as a result, the sum of them may be, to some extent, different from the total score variance. By obtaining the variance components, the researchers used them as the basis for investigating generalizability coefficients (i.e., reliability coefficients) in terms of the theoretical structure of the mean squares for each component (i.e., Kirk's "Expected Mean Square"). Kirk (1982) contended that the expected mean square (EMS) is the value of the mean square that would be gained (see Table 3). Each mean square from the ANOVA was replaced by its corresponding expected mean square equation so as to determine the estimated components of variance.

Table 3
Expected Mean Square Equations for the Two-Facet, Partially Nested p ×(i : t) Design

Sources of Variation Component	*Variance*	*Expected Mean Square*
$\sigma^2 p$		$nint\ \sigma^2 p + nit\ \sigma^2 pt + \sigma^2 pi, pit, e$ Persons (p)
Text types (t)	$\sigma^2 t$	$npni\ \sigma^2 t + np\ \sigma^2 i,it + \sigma^2 pi, pit, e$
p ×t	$\sigma^2 pt$	$ni\ \sigma^2 pt + \sigma^2 pi, pit, e$
i:t	$\sigma^2 i,it$	$np\ \sigma^2 i,it + \sigma^2 pi, pit, e$
pi,pit,e	$\sigma^2 pi, pit, e$	$\sigma^2 pi, pit, e$

np = the number of persons/ applicants

The estimated variance component for the residual is simply the mean square for residual.

The total sum of squares (SSt) for a $p \times (i : t)$ design into sums of squares for persons, nested items, text types, and the residual is:

$$SSt = SSp + SSi,it + SS_{pi\ te.}$$

If we divide the sums of squares (SS) by their respective degrees of freedom (df), it gives the mean squares (MS). According to Kirk (1982), on average, if we repeat the examination of samples from the same population and the universe from the same design, we can achieve the expected mean square (EMS) which is the value of the mean square. Also, according to Shavelson and Webb (1991), the formula for relative and absolute error variances are as follows:

$$\sigma^2_{Rel} = \sigma^2_{Pt} + \sigma^2_{Pi:t} = \frac{\sigma^2_{Pt}}{n'_i} + \frac{\sigma^2_{Pi,pit,e}}{n'_i n'_t}$$

and

$$\sigma^2_{Abs} = \sigma^2_t + \sigma^2_{pt} + \sigma^2_{p:t} + \sigma^2_{Pi:t}$$

$$=$$

$$\frac{\sigma^2_t}{n'_t} + \frac{\sigma^2_{Pt}}{n'_t} + \frac{\sigma^2_{i,it}}{n'_i n'_t} + \frac{\sigma^2_{Pi,pit,e}}{n'_i n'_t}$$

Where facet *i* is confounded with the text-by- item interaction, i.e.,nested in facet *t,* and *n*' indicates the number of conditions of a facet.

4. Results and Discussion

4.1. Results

As statistically found by Rupp et al. (2006), in the case of the construct, reading comprehension has assessment specificity which is basically identified by the design of items and selection of texts. In this line, the frequency and percentage of correct responses for the 10 reading comprehension items are displayed in Table 4. Note that the percentages are equal to the traditional item facility indices. It is clear from Table 4 that item 2 is the easiest item while item 10 is the most difficult. Over 78 percent of the participants gave a correct response to item 2 while less than half of the participants correctly responded to item 10. Traditionally, the item facility indices in the range of .30 to .70 (or from 30 to 70 percent) are acceptable (Green, 2013).

The first five items have facility indices larger than 70 percent. It means that the majority of the students have correctly responded to the items. In other words, these items have been rather easy for all participants taking the test.

Hence, three points emerge from the inspection of the results in Table 4. First, the overall set of items is rather easy for the group of examinees who have taken this test. Second, this easiness will affect item discrimination. The latter is because some items such as item 2 have been correctly answered by the vast majority of the participants. Hence, the examinees are capable of giving a correct response regardless of their ability level.

Table 4

Frequency and Percentage of Correct Responses

Item	Correct	Incorrect	Total	Percentage of correct responses
1	245	76	321	76.3
2	252	69	321	78.5
3	249	72	321	77.6
4	236	85	321	73.5
5	238	83	321	74.1
6	225	96	321	70.1
7	180	141	321	56.1
8	183	138	321	57.0
9	188	133	321	56.6
10	155	164	321	48.9

Finally, it is a well-known statistical fact that the difficulty of items included in a test must match the ability levels of the examinees for the measurement to work properly (Bond & Fox, 2007). That is, if the items are too easy, almost all examinees will give a correct answer. On the other hand, if all items are too difficult, then almost all examinees will fail the items. In both cases, the items will not be able to distinguish between the test-takers. For the test to work properly, the item difficulties must match the ability levels of the examinees. It appears from Table 4 that the difference between the item facilities of the easiest and the most difficult items is less than 30 percent. It means that the item facilities are not widely scattered. Therefore, the items have their best performance over a rather narrow range of abilities.

It is widely known that one of the factors affecting reliability is the amount of variance in the scores (Bachman, 1990). Variance is essentially the deviation of the scores from the mean. That is, the larger the differences between the test scores, the larger the variance. A larger variance will be obtained if the examinees gain widely different scores. A prerequisite for this to happen would be to have items that can easily discriminate among the

examinees. Maximum discrimination is obtained when item facilities are around 0.50 or 50 percent. This is of course correct provided that the item responses are true representations of the abilities of the examinees.

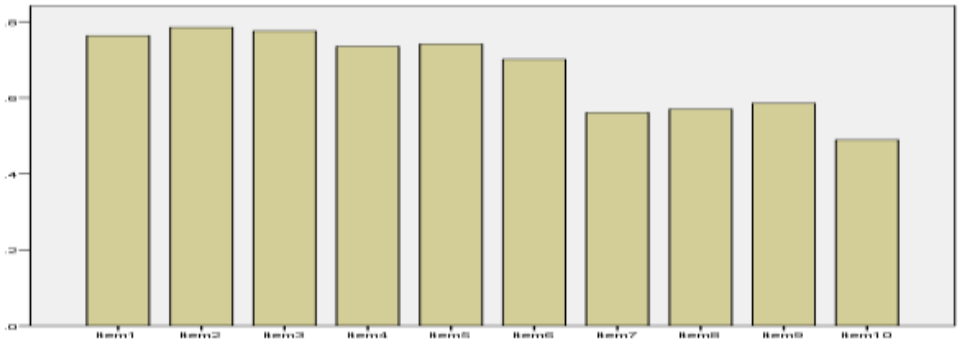


Figure 2. Item Facilities

The item facilities are graphically displayed in Figure 2. It is clear from the table 4 that there is a rough descending order of item facilities from items 1 to 10. Because the items are based on two separate texts, each of which has 5 items, the items in the second text (i.e., items 6 to 10) seem to be more difficult.

The item-total statistics displayed in Table 5 provide further information about the performance of the individual items. The second column shows the mean of the scores if a particular item is deleted. The third column shows the same information for item variance. The corrected item-total correlations in the fourth column show the discrimination indices.

Table 5

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Cronbach's Alpha if Item Deleted
Item1	5.94	3.866	.492	.564
Item2	5.92	3.997	.428	.579
Item3	5.93	4.152	.320	.601
Item4	5.97	3.984	.392	.584
Item5	5.97	3.864	.472	.567
Item6	6.01	4.225	.231	.619
Item7	6.15	4.119	.250	.616
Item8	6.14	4.287	.164	.636
Item9	6.12	4.313	.153	.638
Item10	6.22	4.265	.172	.635

Due to the high-stakes nature of the test, it is expected that only items with excellent performance are included in the test. Hence, we should adopt the criteria for "very good items" from Popham's (2000) guidelines. Therefore, item discrimination indices higher than 0.40 are acceptable. Unfortunately, only three items have discriminations higher than 0.40. It means that 70 percent of the items do not have the required discrimination level. This can seriously jeopardize the validity of the test. Note also that all three items with acceptable discrimination belong to the first text. In addition, three of the items included in the second text have "poor" discriminations. Among them, one reason might be that the examinees do not focus on later item due to time constraints, fatigue, boredom, and so on.

The last column in Table 5 shows the reliability of the test if a particular item is deleted. Cronbach's alpha for the entire set of items was 0.63. The information in Table 5 shows that deleting items 1, 2, and 4 from the test would seriously lower test reliability. Note that these are the items with the highest discrimination value. On the other hand, the omission of the items with the lowest discrimination level slightly improves the reliability. The direct relationship between item discrimination and test reliability is clear here.

All in all, it is clear that the items are not functioning appropriately and this partly accounts for the low reliability of the scores. Based on the guidelines offered by Popham (2000), the ability of 70 percent of the items to discriminate among the applicants is poor. This directly affects test reliability.

4.1.1. Generalizability Theory Analyses

The next step in the analysis of the data was to apply G-theory to answer the research questions. The results of the first round of the analyses are displayed in Table 6. The first column shows the main effects and interactions among the facets. Since the items facet is nested within the text type facet, the main effect for items is confounded with the items by text type interaction effect. In other words, the main effect for items cannot be separated from the interaction between items and text types.

The most important information in Table 6 probably comes in the penultimate column. This column shows the percentage of variance explained by each source of variance or variance component. The largest share of variance is explained by the *pi:t*. This is the interaction between persons and items nested within texts. It is also confounded with all other undetected sources of variance. It explains almost 80 percent of the variance in the data.

The next largest variance belongs to the *persons* facet. This variance component helps to increase the reliability as it is a true variance. A large

variance component for *persons* means that the examinees are not at the same level of ability. All else being equal, the larger the variance component is for persons, the larger the reliability of the test. This facet explains 15.1 percent of the variance in the data.

Table 6

The Variance Explained by each Variance Component

				Components				
Source	SS	Df	MS	Random	Mixed	Corrected	%	SE
P	155.84735	320	0.48702	0.03520	0.03520	0.03520	15.1	0.00398
i:t	33.06916	8	4.13364	0.01230	0.01230	0.01230	5.3	0.00576
T	0.68816	1	0.68816	-0.0021	-0.0021	-0.00211	0.0	0.00120
p× *t*	43.21184	320	0.13504	-0.0101	-0.0101	-0.01019	0.0	0.00237
p× *i:t*	476.13084	2560	0.18599	0.18599	0.18599	0.18599	79.7	0.00520
Total	708.94735	3209					100	

The third largest variance component is the *i:t*. This denotes the main effect of items and the interaction between items and texts. The facility indices in Table 4 showed that items were not at the same level of difficulty. So, the majority of this variance may be due to the differences in item difficulty. This facet explains 5.3 percent of the variance in the data.

Also, the variance component for persons/ Ph.D. applicants is larger than any of the others and the variance component for texts shows systematic differences in the way the applicants responded to the two passages/texts. The relatively large person-by-text interaction indicates that the rank order of the applicants was different on the two passages/texts. This finding indicates that any generalizations about the applicants' relative standing based on either one of the texts by itself would not be dependable and would lead to different conclusions about the applicants' performances in reading comprehension. Moreover, the relatively large three-way, person-by-item-by-text interaction indicates that the observed two-way person-by-text interaction is not the same across the various items. And, the unexplained residual variance is small relative to the other variance components.

The remaining facets (i.e., *t* and *pt*) do not have any contribution. The main effect for texts (i.e., *t*) is zero denoting that both texts are at the same level of difficulty. Similarly, the *pt* effect is zero meaning that there is no interaction between persons and texts.

Now that the relative contribution of each variance component is clear, the G-study can be done. The results are displayed in Table 7. Note that the first two columns pertain to the object of measurement while the rest of the columns are related to the facets. The relative error variance is related to relative or norm-referenced decisions. On the other hand, the absolute error

variance is related to absolute or criterion-referenced decisions (see Shavelson & Web, 1991). Note that only interaction terms affect relative error variance.

Table 7

Contributions of Different Facets to Relative and Absolute Error Variance

Source of Variance	Differentiation Variance	Source of variance	Relative error variance	% relative	Absolute error variance	% absolute
P	0.03520		.....		.....	
	.....	i:t	.....		0.00123	6.2
	.....	t	.....		(0.00000)	0.0
	.....	pt	(0.00000)	0.0	(0.00000)	0.0
	.....	pi:t	0.01860	100.0	0.01860	93.8
Sum of Variances	0.03520		0.01860	100%	0.01983	100%
Standard Deviation	0.18761		Relative SE:	0.13638		Absolute SE: 0.14081
Coef_G relative	0.65					
Coef_G absolute	0.64					

Neither *i:t* nor *t* does affect relative error variance. The amount of variance explained by the interaction between texts and persons is also zero, as was evident from the variance analysis of variance in Table 6. On the other hand, *i:t* explains 6.2 percent of the error variance for absolute decisions. This facet pertains to the differences in item difficulties which do not vary from person to person.

Another important point in Table 7 is related to the amount of absolute and relative error variances. Naturally, the absolute error variance is slightly larger because there are more facets contributing to this variance. Finally, the relative and absolute generalizability coefficients are 0.65 and 0.64, respectively. This may not be acceptable for such a high-stake test. It should be born in mind, however, that this is the reliability of the reading section only. The reliability of the entire test might be much larger.

4.1.2. Decision Studies

In order to find an optimum design for the test, a number of decision studies were run. In order to find the effect of the *items* facet on the generalizability coefficients and the error variances, 10 decision studies were run. All facets and their relevant interactions were kept the same in the decision studies. However, the number of items was changed each time. The number of items ranged from 1 to 10. That is, the first decision study had only 1 item while the last decision study had 10 items.

Table 8

Results of the D-studies

	Relative		Absolute	
Number of items	G coefficient	Error variance	G coefficient	Error variance
1	0.27458	0.09299	0.26201	0.09914
2	0.43085	0.04650	0.41522	0.04957
3	0.53173	0.03100	0.51576	0.03305
4	0.60223	0.02325	0.58680	0.02479
5	0.65428	0.01860	0.63966	0.01983
6	0.69428	0.01550	0.68053	0.01652
7	0.72599	0.01328	0.71307	0.01416
8	0.75174	0.01162	0.73960	0.01239
9	0.77306	0.01033	0.76164	0.01102
10	0.79101	0.00930	0.78023	0.00991

The results of the decision studies are displayed in Table 8. It shows that the generalizability coefficients increase as the number of items increase. As it was pointed out above, the items examined in this study are only a section of the entire set of items included in the test. Hence, no hard and fast rules can be offered for the optimum reliability for the reading section of the test. It is clear from the table 8 that both relative and absolute G-coefficients are very close. This is because there was only one extra facet contributing to absolute error variance.

The absolute and relative error variances are also reported in Table 8. Again, by considering this table, we can understand that the two sets of variances are very close. In fact, there is little difference between relative and absolute error variances. It should also be pointed out that the error variances decrease as the number of items increases.

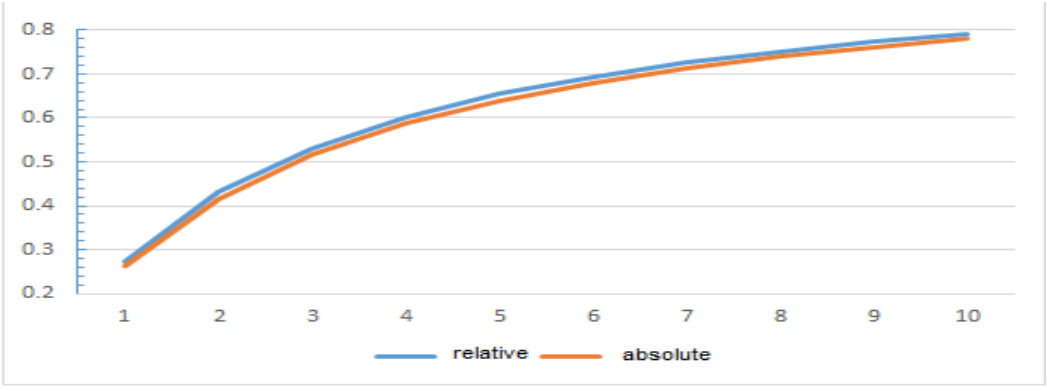


Figure 3. Relative and Absolute G-coefficients

The results of the Decision (D-) studies are also graphically presented in Figure 3 and Figure 4. Figure 3 shows the G-coefficients while Figure 4 displays the relative and absolute error variances. It appears that with increasing number of items in the text types, the relative and absolute error variances decrease and this reduction of the error variances leads to an increase in relative and absolute G-coefficients. This finding is in agreement with previous research (e.g., Brennan et al., 1995; Swartz et al. 1999; Schoonen, 2005; Lee & Kantor, 2007; Gebril, 2010). Therefore, this result of the study is consistent and in line with five previous studies in which reduction of the error variances leads to an increase in relative and absolute G-coefficients.

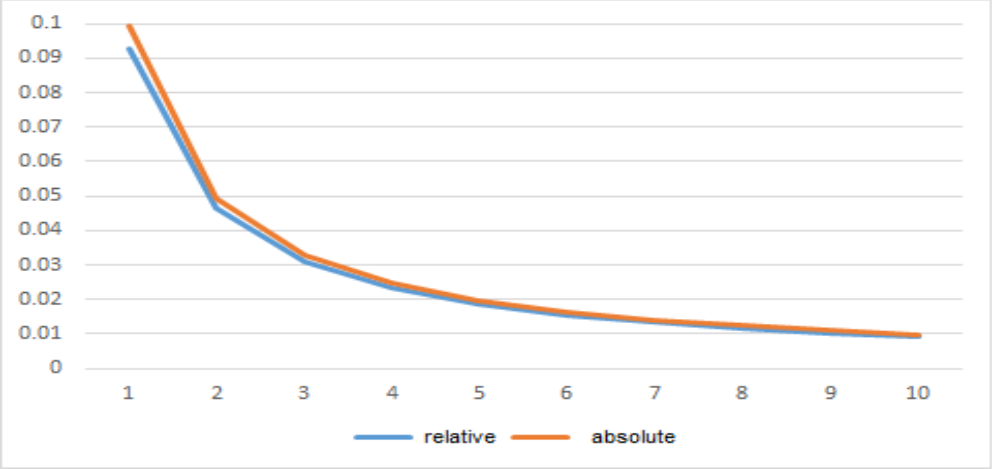


Figure 4. Relative and Absolute Error Variances

4.2. Discussion

Brown & Glasner (1999) concluded that in order to increase the quality of educational systems, assessment should play the function in higher education process. Therefore, more attention has been paid to the academic standards in terms of the relation between the students' entry level and the outcomes of the assessment. Nevertheless, as elaborated on by van de Watering and van der Rijt (2006), "little is known about the degree to which assessments in higher education are correctly aimed at the students' levels of competence" (p. 134).

By referring to Bachman's (1990) claim, the deviation of the scores from the mean impacted the reliability. The largest variance of the *persons* indicates they obtained widely different scores, and they were of different ability levels. This finding shows the systematic differences in the applicants' responses in the texts. Moreover, it is in line with earlier research, showing that persons widely vary in their performance on responding to reading

comprehension texts (e.g., Brennan, 2010; Cardinet et al, 2009; Fan & Sun, 2014; Shavelson & Webb, 1991). In addition, the *items* facet had a noticeable amount of variance, and therefore, in the decision studies the focus was on the impact of items; it clarifies the main effect of the items and their interaction with texts. This finding is consistent with the findings of some researchers who maintained that the items types are among the possible causes of variation in the test takers' performances in the reading comprehension, and thus this result is in line with the findings of some scholars (e.g. Mostow et al., 2016; Schoonen, 2005; Steensel et al., 2012; Vidal-Abarca et al., 1998;) whose findings provide some positive evidence in support of the different performance of examinees based on the type of items. However, until now, there is little research on the persons, items, genres and their interaction based on g-theory, and most of the analyses were almost always based on single task within different text types (Barkaoui, 2007; In'nami & Koizumi, 2015; Lee ,2005; Stansfield et al., 1992b; Wang, 2010) or based on the impact of text types on the generalizability of scores (Bouwer et al. 2015; Gebril, 2010; Han & Ege, 2013; Huang & Han, 2013; Huang & Foote, 2010; Schoonen, 2005; Van den Bergh et al., 2012). Therefore, the effects of items, and text types were confounded and, consequently, deductions that could be drawn about systematic differences within and across text types were limited. The current study expands this knowledge by untangling the effects of items and genre by including persons, items, and text types and unmeasured variation in the measurement. The results show that items have an effect above and beyond specific genre effects. Furthermore, the relatively large person-by-text interaction shows that the rank order of the applicants was different on the two texts. According to Brennan, Goa, and Colton (1995), this substantially large value can be based on the rank ordering of text difficulty which is different for the different applicants, or that the rank ordering of applicants differs by text type to a notable degree. This finding shows the effect persons had on the texts in their interaction, and concurs with those of the earlier investigations reviewed above. In other words, the general trend of a larger person-by-text interaction effect found in this research is similar to that of the previous studies. Hence, this result is also supported by the findings some researchers (e.g., Brennan, et al., 1995; Brown, 2011; Huang, 2009; In'nami & Koizumi, 2015; Lee & Kantor, 2007; Schoonen, 2005). This large value of variance is also common in writing tests (Gebril, 2010).

On the other hand, the G-study showed that the main effect for the *text types* facet was zero. It means that both texts were at the same level of difficulty, and they did not have any contribution in the variance. This finding means that the text types did not impact on the applicants' performances in reading comprehension. This finding is in contrast to some other researchers' findings that text type or genre is one possible source of

variation, and the generalizability of scores differs from one text type to another, and therefore, the effects of text types contaminate persons' performance. For example, Bouwer et al. (2015) in their study found that the generalizability of scores differs from one task type to another. Hence, this finding does not support those researchers' findings (e.g., Bouwer et al., 2015; Lee, 2005; Schoonen, 2005; Van den Bergh et al., 2012). Therefore, more studies should be conducted to consider this issue. In addition, the interaction between persons, items, and texts, i.e., the residual shows that the observed two-way person-by-text interaction is not the same across the various items, and the magnitude of its variance over the other variance components was small. This finding is also supported by some scholars' findings (e.g., Brennan, 2001; Shavelson & Webb, 1991). This effect shows the random error source. For instance, it may come out when a person breaks his or her pencil during the examination and loses time on later items, causing his or her score to be lower than it should be.

The findings of this study provide important implications for foreign language assessment and supports new trends towards preparing and assessing reading comprehension. In addition, the intact items were mostly in the argumentative text which is the second text in the reading comprehension section. Therefore, current results suggest that considerations should be taken about these issues for these entrance exams, and Ph.D. applicants whose majors are not English Language. Certainly, this movement can lead to positive washback in this kind of assessment. Moreover, the results of this study also suggest that the item types play a very important role in the measurement precision of the applicants' ability in reading comprehension. As a result, this study informs the researchers and those who are involved in preparing, designing and evaluation of reading comprehension about the magnitude of the types of errors, so that decisions concerning whether error magnitudes are within acceptable ranges can be applied to future studies. Hence, a desired level of generalizability can be obtained in those studies.

5. Conclusion and Implications

As mentioned before, generalizability theory (G-theory) "has been hailed as a liberalization of classical test theory" (Vispoel, et al. 2018, p. 2). Some parts of the findings of this study are consistent with other scholars' findings, but some other parts are not. However, a number of implications can come from this study. As Fan and Sun (2014) argued that the G-study acts as a “pilot” reliability study, this research can supply information for planning the “real” test. Yet, by referring to Lee and Kantor's (2007) claim, the power of G-theory analyses is based on a large sample of population for each facet in the universe of permissible observations.

The findings of this study provide important implications for foreign language assessment and support new trends towards preparing and assessing reading comprehension. In addition, the intact items were mostly in the argumentative text which is the second text in the reading comprehension section. Therefore, current results suggest that considerations should be taken about these issues for these entrance exams, and Ph.D. applicants whose majors are not English Language. Certainly, this movement can lead to positive washback in this kind of assessment. Moreover, the results of this study also suggest that the item types play a very important role in the measurement precision of the applicants' ability in reading comprehension. As a result, this study informs the researchers and those who are involved in preparing, designing and evaluation of reading comprehension about the magnitude of the types of errors, so that decisions concerning whether error magnitudes are within acceptable ranges can be applied to future studies. Hence, a desired level of generalizability can be obtained in those studies.

References

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

Bachman, L. F. (2004). *Statistical analyses for language assessment book*. Cambridge University Press.

Bachman, L. F. (2010). *Fundamental considerations in language testing*. Oxford: Oxford University Press.

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

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86 –107.

Berry, V. (1993). Personality characteristics as a potential source of language test bias. *Language testing: New openings. Jyvaskyla, Finland: Institute for Educational research. University of Jyvaskyla.*

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: fundamental measurement in the human sciences.* New York and London: Routledge.

Bouwer, R., Béguin, A., Sanders, T., & Van den Bergh, H. (2015). Effect of genre on the generalizability of writing scores. *Language Testing*, *32*(1), 83-100.

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1-21.

Brennan, R. L. (2001a). *Generalizability theory*. New York: Springer Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA*. Iowa City, IA: Iowa Testing Programs, University of Iowa.

Brennan, R. L. (2000). Performance assessment from the perspective of generalizability theory. *Applied Psychological Measurement, 24*(4), 339–353.

Brennan, R., Goa, X., & Colton, D. (1995). Generalizability analyses of Work Keys Listening and Writing tests. *Educational and Psychological Measurement*, *55*(2), 157–176.

Brennan, R. L. (1992). Elements of generalizability theory. Iowa City, IA.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement*, *14*(3), 277-289.

Brown, J. D. (2011). What do the L2 generalizability studies tell us? *International Journal of Assessment and Evaluation in Education, 1*, 1–37.

Brown Jr, J. and Glasner, A. (1999). *Assessment matters in higher education*. McGraw-Hill Education: UK.

Burt, C. (1936). The analysis of examination marks. In P. Hartog & E. C. Rhodes (Eds.), *The marks of examiners* (pp. 245-314). London: Macmillan.

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG* Taylor & Francis.

Cardinet, J., Tourneur, Y., & Allal, L. (1981). Extensions of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement, 18,* 183–204.

Cardinet, J., Tourneur, Y., & Allal, L. (1976). The symmetry of generalizability theory: Application to educational measurement. *Journal of Educational Measurement, 13*, 119–135.

Cerdan, R., Vidal-Arbarca, E, Martinez, T., & Gil, L. (2009). Impact of question-answering tasks on search processes and reading comprehension. *Learning and Instruction, 19*(1),13–27.

Crick, J. E. (1983). Manual for GENOVA: a generalized analysis of variance system, Iowa. *American College Testing Program*.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and psychological measurement*, *64*(3), 391-418.

Cronbach, L. J., Nageswari, R., & Gleser, G. C. (1963). Theory of generalizability: Aliberation of reliability theory. *The British Journal of Statistical Psychology, 16*, 137–163.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements*. New York: Wiley.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16*, 407– 424.

Eason, S. H. (1989). Why Generalizability Theory Yields Better Results than Classical Test Theory.

Everitt, B. S., and Howell, D.C. (2005). Repeated measures analysis of variance. *Encyclopedia of Statistics in Behavioral Science*.

Fan, C., H. and Hansmann, P., R. (2015). Applying generalizability theory for making quantitative RTI progress-monitoring decisions. *Assessment for Effective Intervention,40(4),* 205–215.

Fan, X., and Sun, S. (2014). Generalizability theory as a unifying framework of measurement reliability in adolescent research. *Journal of Early Adolescence. 34(1)*, 38–65.

Feldt, L. S., & Brennan, R. L. (1989). *Reliability*. In R. L. Linn (Ed.), Educational measurement (pp. 105–146). Washington, DC: The American Council on Education/Macmillan.

Ferrara, S. (1993, April). Generalizability theory and scaling: Their roles in writing assessment and implications for performance assessments in other content areas. In *annual meeting of the National Council on Measurement in Education, Atlanta*.

Finlayson, D. S. (1951). The reliability of the marking of essays. *British Journal of Educational Psychology*, *21*(2), 126-134.

Gebril, A. (2010). Bringing Reading-to-Write and Writing-Only Assessment Tasks Together: A Generalizability Analysis. *Assessing Writing, 15*(2), 100–117.

Gleser, G. C., Cronbach, L. J., & Rajaratnam, N. (1965). Generalizability of scores influenced by multiple sources of variance. *Psychometrika, 30*, 395–418.

Green, R. (2013). *Statistical analyses for language testers*. New York: Palgrave Macmillan.

Han, T., & Ege, İ. (2013). Using generalizability theory to examine classroom instructors' analytic evaluation of EFL writing. *International Journal of Education*, *5*(3), 20.

Hoyt, C. J. (1941). Test reliability estimated by analysis of variance. *Psychometrika, 6*, 153–160.

Huang, C. (2009). Magnitude of task-sampling variability in performance assessment: A metaanalysis. *Educational and Psychological Measurement, 69*(6), 887–912.

Huang, J. (2012). Using generalizability theory to examine the accuracy and validity of large scale ESL writing assessment. *Assessing Writing, 17*, 123–139.

Huang, J. (2008). How accurate are ESL students' holistic writing scores on large-scale assessments? A generalizability theory approach. *Assessing Writing, 13*, 201–218.

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*, 219–333.

Huang, J., & Han, T. (2013). Holistic or analytic – A Dilemma for Professors to Score EFL Essays? *Leadership and Policy Quarterly, 2*(1), 1–18.

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.

Kane, M. (2010). Validity and fairness. *Language testing*, *27*(2), 177-182.

Kendeou, P., McMaster, K. L., & Christ, T. J. (2016). Reading comprehension: Core components and processes. *Policy Insights from the Behavioral and Brain Sciences*, *3*(1), 62-69.

Kieffer, K. M. (1998). *Why Generalizability Theory is Essential and Classical Test Theory is Often Inadequate?* [Proceeding]. *Paper Presented at the Annual Meeting of the SouthWestern Psychological Association*. New Orleans, LA: USA.

Kirk, R. E. (1982). *Experimental Design* (2nd edition). Belmon. CA: Brooks/ Cole.

Kunnan, A. J. (1992). An investigation of a criterion-referenced test using G-theory, and factor and cluster analyses. *Language Testing*, *9*(1), 30-49.

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.

Lee, Y. W., & Kantor, R. (2007). Evaluating prototype tasks and alternative rating schemes for a new ESL writing test through G-theory. *International Journal of Testing, 7*, 353–385.

Lin, C. K. & Zhang, J. (2014). Investigating correspondence between language proficiency standards and academic content standards: A generalizability theory study. *Language Testing*, 7, 1–19.

Linn, R. L. (1981) *'Curricular validity: Convincing the courts that it was taught without* precluding the possibility of measuring it', the Ford Foundation, Boston College, MA.

Lord, F. M. (1959). Test of the same length do have the same standard errors of measurement? *Educational and Psychological Measurement, 19*, 233–239.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325–336.

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158–180.

Welch, R. C. (2014). *Action research from concept to presentation: A practical handbook to writing your master's thesis*. Author House.

Marcoulides, G. A. (2000, March). *Generalizability theory*: Advancements and implementations. Invited colloquium presented at the 22nd Language Testing Research Colloquium, Vancouver, BC, Canada.

Marshall, E. (1998). The Marshall plan for novel writing. Cincinnati, OH: Writer's Digest Books.

Messick, S. (1989). Validity In. R. Linn (Ed.) *Educational measurement* (pp.13-103).

Morrell, J. (2006). *Between the lines: Master the subtle elements of fiction writing*. Media, Inc.

Mostow, J., Huang, Y. T., Jang, H., Weinstein, A., Valeri, J., & Gates, D. (2017). Developing, evaluating, and refining an automatic generator of diagnostic multiple choice cloze questions to assess children's comprehension while reading. *Natural Language* Engineering, *23*(2), 245-294.

Pearson, P., Valencia, W., & Wixson, K. (2014). Complicating the world of reading assessment: Toward better assessments for better teaching. *Theory into Practice, 53*(3), 236–246.

Popham, W. J. (2000). *Modern educational measurement*. Boston, Allyn & Bacon.

Rouet, J. F., Vidal-Abarca, E., Erboul, A. B., & Millogo, V. (2001). Effects of information search tasks on the comprehension of instructional text. *Discourse Processes*, *31*(2), 163-186.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language testing*, *23*(4), 441-474.

Selgin, P. (2007). *By cunning & craft: Sound advice and practical wisdom for fiction writers*. Writer's Digest Books.

Shavelson, R.J., & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbury Park, CA: Sage.

Smith, Jr., E. V. & Kulikowich, J. M. (2004). An application of generalizability theory and many facet rasch measurement using a complex problem solving skills assessment. *Educational and Psychological Measurement, 64,* 617–639.

Stansfield, C. W., & Kenyon, D. M. (1992). Research on the comparability of the oral proficiency interview and the simulated oral proficiency interview. *System*, *20*(3), 347-364.

van de Watering, G., & van der Rijt, J. (2006). Teachers' and students' perceptions of assessments: A review and a study into the ability and accuracy of estimating the difficulty levels of assessment items. *Educational Research Review*, *1*(2), 133-147.

van Steensel, R., Oostdam, R., & van Gelderen, A. (2013). Assessing reading comprehension in adolescent low achievers: Subskills identification and task specificity. *Language testing*, *30*(1), 3-21.

Strube, M. J. (2002). *Reliability and generalizability theory*. In L.G. grimm & P.R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Sudweeks, R.R., Reeve. S., Bradshaw, W. S. (2005). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing 9*, 239–261.

Swartz, C.W., Hooper, S.R., Montgomery, J.W., Wakely, M.B., De Kruif, R.E.L., Reed, M., Brown, T.T., Levine, M.D., & White, K.P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement, 59,* 492–506.

Thompson, B. (2003). *A brief introduction to generalizability theory*. In B. Thompson (Ed), Score reliability: Contemporary thinking on reliability issues (pp. 43– 58). Thousand Oaks, CA: Sage.

Van den Bergh, H., De Maeyer, S., Van Weijen, D., & Tillema, M. (2012). Generalizability of text quality scores. *Measuring writing: Recent insights into theory, methodology and practices*, *27*, 23-32.

Vidal-Abarca, E., Gilabert, R., & Rouet, J. F. (1998). The role of question type on learning form scientific text. *Paper presented at Seminario 'Comprension y produccion de textos cientificos'*, Aveiro, Portugal.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *Journal of personality assessment*, *100*(1), 53-67.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2017). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1.

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2016). Using G-theory to enhance evidence of reliability and validity for common uses of the Paulhus Deception Scales. *Assessment*, *25*(1), 69-83.

Wang, H. (2010). Investigating the justifiability of an additional test use: An application of assessment use argument to an English as a foreign language test (Doctoral dissertation). Retrieved from ProQuest. (AAT 3441468)

Webb, N.M., Rowley, G. L., & Shavelson, R. J. (1988). Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development, 21*, 81– 90.

Webb, N. M., & Shavelson, R. J. (2005). Generalizability theory: Overview. *Encyclopedia of statistics in behavioral science*.

Wu, Y-F. and Tzou, H. (2015). A Multivariate generalizability theory approach to standard setting. *Applied Psychological Measurement. 39(7),* 507–524.

Zhang, J., & Lin, C. K. (2016). Generalizability theory with one-facet nonadditive models. *Applied psychological measurement*, *40*(6), 367-386.

Bibliographic information of this paper for citing:

Ahmadi Shirazi, M., Alavi, S. M., & Salarian, H. (2019). An investigation into item types and text types of reading comprehension section of Iranian Ph.D. entrance exams using G-theory. *Journal of Modern Research in English Language Studies, 6*(1), 1-29.