Journal of Education & Human Development
March 2014, Vol. 3, No. 1, pp. 327-346
ISSN: 2334-296X (Print), 2334-2978 (Online)
Copyright © The Author(s). 2014. All Rights Reserved.
American Research Institute for Policy Development
42 Monticello Street, New York, NY 12701, USA.
Phone: 1.347.757.4901 Website: www.aripd.org/jehd

Threats to Validity: construct-irrelevant variances contributing to performance under-representation on Graduate Record Exam (GRE)

Ali Panahi[*]

Abstract

The fundamental rationale for advancing the present investigation is traced back to the recognition that construct-irrelevant variances and construct under-representation threaten the value implications and the consequential validity of the tests, hence spoiling the appropriateness and usefulness of the test scores for the purpose of decision-making (Mesick, 1989). Accordingly, this article deals in the main with the potentially interfering variances giving rise to performance under-representation on Graduate Record Exam. To these tight objectives, after undertaking the pilot testing project, dividing the subjects into two groups and sketching out the instruments, the run *t-test* as well as the self-reported protocols ultimately revealed that some statistically significant differences existed in both anxiety and performance of the two groups, i.e., math less-interested-in test-takers and math more-interested-in ones as a function of anxiety and math-section-related face validity, i.e., test bias. Also, the verbal section of GRE aiming at measuring the intelligence of the test takers is argued. Pedagogically assumed, the concentration is placed on the premise that the extraneous and interfering parameters culminate in misinterpretation and misuse of the tests in which confidence is put so as to make some global or local decision. Hence, they must be to the feasible extent controlled and manipulated so as to enhance the validation of and eschew misconstruing and misapplying the intended operationalized trait and score meaning for judgmental undertakings and social enterprises.

Keywords: Construct; Trait; Construct- Irrelevant Variances; Construct Under-representation; Performance under-representation; Consequential Validity; Validation

[*] PhD Candidate in TEFL, University of Tehran, Iran

1. Introduction

At a recent academic conference, The 1st National Conference on Emerging Horizons in ELT and Literature at Islamic Azad University, Ahar Branch, Iran, I, as an author of this paper, began to put the vague academic perspectives I had into an argumentative framework with respect to GRE, expecting that the debate will lead its quarrelsome academic life. As such, as an academic, the personal belief of mine stands on the footing that no test does exist without taking the construct-irrelevant variances and construct-dependent factors giving rise to validity discussion in terms, in the main, of the consequences and social values of the test for decision making ambitions into an attentive account. The clear-cut rule is that GRE has a kind of intelligence-mediated design, the score meaning of which confidence is put in to make decision. Hence, validity and validation issues related to GRE should be born in mind as with other global or local tests. Thus, the discussion of anxiety and face-validity-related issues as construct irrelevant variances threatening performance on GRE come to the front.

To the present, however, some research findings have underscored the key preventive actions concerning the impact anxiety, leading to performance under representation, exerts on the test takers. Given these tenet, the relationship between anxiety and performance has been of interest for decades (Zhang, 2013), but this has not been hotly tackled in the field of GRE. Clearly sufficient, identifying the anxiety source can help the test takers alleviate it and lead to better performance; of course, the debate concerning the relationship between anxiety with regard to causation or consequence is has long been in evidence. (Zhang, 2013). Also, since the face of discussion is with validity and construct irrelevant variances threatening it, it is required to be in brief uttered that face validity can be a superficial factor leading to under-performance or over-performance.

With a view to the review of literature associated with the definitions and comments exerted on face validity, Brown (2004) states that face validity is the degree to which a test looks right, and appears to measure the abilities it claims to measure. He believes that face validity is occasionally considered as a superficial factor which is dependent on the whim of the perceiver. Contrarily, some rejected it as representative of any type of legitimate and scientific validity evidence, although the fact that the appearance of the assessment may be an important characteristic.

In this regard, Bachman(1995) states that there continues to be considerable misunderstanding concerning the term "face validity". Contrary to these all comments, Henning (1987) utters that face validity is a substantive impression of the extent to which the test fulfills the intended purpose of measurement, stating also that some authors do not distinguish between face validity and content validity. In contradiction, Cronbach holds that adopting a test just because it appears reasonable is a bad practice (as cited in Bachman, 1995, P. 286). To practically deal with the issue, first of all, a brief history of testing gradually being triggered to *construct validation* and validity issues as a central concern of language testing research is outlined below.

2. A Brief history of testing methodology approaches

As far back as written history takes us, language testing, as a sub-field within applied linguistics (Davis,1990), has undertaken some dramatic changes growing conducive to an argumentative framework by researchers and testers. Obviously, symbolic of the changing wind and shifting sand in testing methodology field is the swinging of pendulum from testing to assessment. In this regard, Gipps(1994) takes possession of paradigm shift- a set of interrelated concepts or scientific revolution- which is indicative of beyond testing, as with beyond method in language teaching. The former has in the main the implication of moving from testing to education, from single approach for assessment to triangulation and use of a variety of approaches for operationalizing and measuring a construct, and ultimately, it typifies a move from psychometric testing to educational assessment.

Accordingly, given the consequential and evidential basis (as cited in Bachman,1995), assessment takes on a broad range of purposes, central to which is supporting teaching and learning (Gipps, 1994). In the same vein, Bachman (1999) addresses that in the past twenty years, language testing research and practice have witnessed the refinement and blossoming of a rich variety of approaches and tools for research and development, thus, making a contribution to broadening our horizon of the factors and processes affecting performance on the test, as well as of the consequences and uses of the test (Bachman, 1999). To obtain a brief perception of the approaches, Spolsky (1978) and Hinofotis (1981) pursue that language testing can be broken into three periods of development including the pre-scientific period, the psychometric/structuralist period, and the integrative [psycholinguistic, Weir, 1990]-sociolinguistic period, as appear below.

2.1. Pre-scientific Movement

Language testing has its roots in pre-scientific stage in which no special skill, theory or expertise in testing is required, the one featured by lack of concern for statistical considerations or for such notions as objectivity and reliability; thus, the focus of trend is on the subjective judgment rather than objective judgment of scores (Heaton 1988, Weir 1990; Brown, 1996). Clearly, Hinofotis's (1981) cogent position follows that the pre-scientific movement ended with the onset of the psychometric structuralist movement, signaling the advent of scientific and theory-loaded movement coming up below.

2.2. Psychometric Structuralist Movement

With the onset of the psychometric-structuralist movement of language testing, born in 1970s, language tests, i.e., standard tests such as multiple-choice tests on which the language elements were separated and tested, became increasingly scientific, reliable, objective, theory-supported and precise and were also established on statistical analyses for the first time (Weir,1990; Brown, 1996;Carrol, 1972; Weir, 1990). This approach was criticized on the ground that it took just linguistic competence to the negligence of communicative competence into account. However criticized globally, the psychometric-structuralist tests are still very much in practice around the world and have been supplemented by what Carrol (1972) and Oller (1979) called integrative tests.

2.3. Psycholinguistic- Sociolinguistic Movement

In the early 1980s, the shift of attention from discrete-point tests to global tests (integrative tests) gave rise to psycholinguistic-sociolinguitic movement. The experts and researchers came to believe that language is more than the sum of the discrete elements being tested during the psychometric-structuralist movement (Brown, 1996; Heaton 1991; Oller, 1979; Canale, 1984; Weir, 1990). The criticism came largely from Oller (1979) with the appearance of “unitary trait hypothesis" or " unitary competence hypothesis" associated with a general factor, i.e., G-factor , according to which competence is a unified set of interacting abilities - vocabulary, grammar, phonology, and the four language skills Oller (1983; 1986) - that cannot be tested apart and measured adequately, stating that the *pragmatic test*s, such as cloze tests (gap-fill), dictation tests, the oral interview, and composition writing, require the learner to relate sequences of linguistic elements via pragmatic mappings to extra-linguistic context".

Taken as a whole, the research upon which Oller's claims for a unitary competence were based was eventually rejected (Farhady, 1983). To the testers and researchers' astonishment, Oiler himself (1983) admitted that the unitary competence hypothesis was wrong.

By the mid-1980s, inspired by the work of sociolinguists like Hymes (1967) and Canale and Swain (1980), and Savignon(1983) who assumed the language ability as multi-componential and dynamic rather than static, sociolinguistic and communicative facet of language use and negotiation of meaning was aptly stressed. As Bachman and Palmer (2000) viewed, the relationship between test situation and non-test situation is indicative of performance test, task-based language testing, democratic testing, cooperative testing, etc**.,** which were substantiated for communication requirements. Contrary to the disconfirmation of Oller's hypothesis, it has had a major and lasting impact on the field so that his work established *construct validation* and validity issues as a central concern of language testing research, the one being as a main concern for the present investigation and also the one on which an explanatory attempt is made below.

3. From the birth of validity to the present

For more than 100 years, as it is evident in the categories below, divergent positions on the concept of validity, assuming a pivotal role, have been on the table so that it has been subject to a vast variety of comments and definitions by the researchers and testers in the field; even today, to our astonishment, the meaning of validity is contested. Categorically, to Newton and Shaw (2013), the history of validity is divided into five broad periods:

1. The mid-1800s and 1920: *gestation*
2. 1921 and 1951: *crystallization*
3. 1952 and 1974: *fragmentation*
4. 1975 and 1999: *reunification*
5. 2000 and 2012: *deconstruction*

Explicitly viewed, the opinion runs that the first period is enigmatic of the vague and unclear conceptualization of the notion of validity, when it was a fetus, i.e., before its birth. Thus, there was no fixed, clear and believable definition in existence for validity.

In other words, the researchers in this period presented no outline for a precise conceptualization and definition of validity. However, they expounded their perspectives and comments on the subject at great length until the fetus of validity was born and came into a concrete being, followed by which the second period came to discussion. Vividly, this period, was skewed towards crystallization, during which the concept of validity got an explicit identity and moved beyond the vague notion sketched out in stage one so that the scene for a definitional account of validity was set and delineated. At the same time, when validity started getting more explicit than before, Kelly(1927) demonstrated that " the question [of validity] is thoroughly roused from a slumber of centuries, probably never to sleep again". He, then, in 1927, stated that validity deals with whether a test really measures what it purports to measure (as cited in Schouwstra, 2000). In keeping with this, Birjandi and Mosallanejad (2010, p. 228) go on to argue that the traditional stance for validity is summarized in three conceptualizations: 1. Validity is a property of tests, rather than of test score interpretation. 2. In order to be valid, test scores should measure some purported construct directly 3. Score validity is a function of whatever construct is intended to measure.

But, presently, the definition assumed in stage two was flawed and has been on the whole debated and continues to be more argumentative because what we tend to know together with its extent is not always manageable. Followed by this, stage three signaled, on a general consensus, the genesis of standards tending to focus on types of validity (Newton and Shaw, 2013) including content validity, predictive validity, concurrent validity, construct validity (Messick in 1989, as cited in Fulcher and Davidson, 2007).These four validities came to embody the fragmented view of validity and validation, thus called fragmentation period. The diversity of ideas on validity and validation, i.e., the degree of validity, during this stage introduced a new challenge to the test developers and test designers.

During the years 1975 and 1999, Messick's account of validity and validation emphasized that all of these five validities should be unified as one validity called construct validity, believing that construct validity- as a scientific inquiry into score meaning- ought to be a foundation for all validation-related elements (Newton and Shaw, 2013). Hence, the criticism arose from Messick's (1989, 1996a, 1996b) views that the traditional conception of validity is fragmented and incomplete because it fails to take into account both evidence of the value implications of score meaning as a basis for action and the social consequences of score use.

His alternative approach views validity as a unified concept which places a heavier emphasis on how a test is used and also considers different types of validity evidence for construct validity. Hence, his definition of validity is: an integrated, rather than fragmented, evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy, appropriateness, meaningfulness and usefulness of the inferences and actions based on test scores(as cited in Bachman,1995; Fulcher and Davidson, 2007). Thus, *reunification period* came into existence. The following table is more supportive to the point.

Facets of validity (Messick, 1989a; 1989b)

	Test interpretation	Test use
Evidential basis	Construct validity	Construct validity + relevance/ utility
Consequential basis	value implications	Social consequences

Obviously enough, this model cements the consensus that construct validity is the one unifying conception of validity, and it extends the validity beyond test score meaning to include relevance and utility, value implications, and social consequences. he integrates use, values and consequences. As a matter of effect, the interpretation and use facets of testing (also called "function or outcome" of testing) have other four clear-cut validity categories. Carefully considered, the evidential basis of both test interpretation and test use is construct validity; here, the meaning of the score can be supported with use of evidence coming from different sources, the contextual factors considered. By the same token, Fulcher and Davidson (2007) maintain firmly that central to any validity is the attentive consideration of the intended meaning as well as interpretation and use of test scores.

The consequential basis of the test is the social consequences and real application of the tests; as Bachman (1995) rightly put, validity is associated with "use of test" and reliability, which is not our discussion however closely and vitally related, is connected to " score of test". Attentively regarded, the validity move is from dichotomization to combination, from fragmentation period to reunification period, from an old view, i.e., separation of content validity, construct validity, and criterion-related validity, all of which were separately considered, to modern perspectives.

More supportively, Bachman and Palmer (2000) consider the notion of usefulness, instead of separation of validity into three categories, as an alternative to validity; to them, construct validity includes the combination of reliability, i.e., the consistency of test scores across facets of the test, authenticity, i.e., the relationship between task characteristics and the characteristics of task in real world, interactiveness, i.e., engagement of the test taker's characteristics, such as language ability, background knowledge, motivation,…, in task, and practicality, i.e., test implementation. They go on to state that the overall usefulness (Bachman and Palmer, 2000) changes according to context (Fulcher and Davidson, 2007).

Brown (2004) holds the status, on a general consensus, that validity is the extent to which inferences made from assessment results are appropriate, meaningful, and useful in terms of the purpose of the assessment. He views that validity is established with use of various pieces of evidence, rather than a single source, such as consequential-validity-related evidence, face-validity-related evidence, construct-related evidence, criterion-related validity and content-related evidence, etc.

Axiomatically, during the 1990s, the validity and validation issues were essentially prevailed by unified view of validity. But, with the turn of millennium, new cracks marking the *deconstruction period* began to emerge so that Cronbach and Meehl together with Kane, in 1992 raised some controversies and confusion with respect to Messick's perspectives. On one the hand, it was unclear how to translate construct validity theory into validation practice and it was unclear whether construct validity was the best way to unify validity theory. So, this period was argument-established being concerned with debating the nature and significance of construct validity and that whether the construct validity should be the foundation of all validly Messick had argued. Furthermore, it was debated with regard to the fact that either construct should be considered in terms of measurement and that either Messick considers ethical issues in the interpretation of score meaning or not. All these debates are settled down by Newton and Shaw (2013) who aptly defended Messick's perspectives. The 21st century evaluation is well outlined by Newton and Shaw (2013) who hold the view that the validity is universally regarded as the hallmark of quality for educational measurement.

4. Definition of construct

When the face of discussion comes to the word "construct", we do think of abstract nouns [or psychological concepts to me], such as love, intelligence, anxiety, etc.

Construct is characterized by being measurable or operationalizable through getting linked to something observable (Fulcher & Davidson, 2007). A commonly understandable definition in testing field is that construct is what we would like to test; so, it is synonymous with trait, abilities, performance, and characteristics of the testees we would like to test (Fulcher & Davidson, 2007; Bachman, 1995). By this implication, a construct is any theory, hypothesis or model that attempts to explain an observed phenomenon. It asks: does this test tap into the theoretical construct as it has been defined. To put the point on a more technical footing, tests are operational definitions of constructs, in that they operationalize the entity that is being measured (Fulcher & Davidson, 2007).

Given a more technically detailed elaboration, Backman(1999) illustrate a dialectic between what has been called "trait/ability-focused" and "task/context-focused" perspectives on or approaches to defining constructs in language testing. Bachman states that from the early 1960s to the present, we can see a dialectic between a focus on *language ability/trait* as the construct of interest and a focus on *task* or *context* as the construct. Likewise, (Chapelle, 1998, as cited in Fucher & Davidson, 2007) elaborates on three different approaches to construct definition: trait, behaviorist, and interactionalist models. The trait model views construct as context-free and something fixed and unchangeable.

In behaviorists' model, construct is affected by facets or contextual factors, such as physical setting, topic, and participants. The view states that our ability or construct changes from situation to situation and it is not a stable competence. Thus, the difference between these two perspectives is implicit in the generalizability of the score meaning from one situation to another situation. With regard to interactionalist view of construct, Bachman(1999) says that more recently, some researchers, drawing largely on research in social interaction and discourse analysis, have proposed an interactionalist perspective on language assessment, which views the construct we assess not as an attribute of either the individual language users or of the context, but as jointly co-constructed and residing in the interactions that constitute language use. Bachman pursues to state the three general approaches for defining the construct, or what we want to assess: 1) ability-focused, *2)* task-focused, and 3) interaction-focused, believing that these three are not mutually exclusive, but inclusive.

Of particular relevance here is their application of Chapelle's1998 (as cited in Fulcher & Davidson, 2007) definition of the interactionalist approach, indicating that an interactionalist approach to inferences from construct requires to be triggered to particular contexts. Likewise, Bachman (1999) described what he called the "fundamental dilemma associated with interwoven relationship between context and construct; these all are illustrative of the difficulty of the operational definition of construct.

5. Construct-irrelevant variances (CIV): Factors threatening validity

The most obvious threat to validity arising from the misapplication of tests or misinterpretation of score meaning (Henning, 1987) is construct-irrelevant variance (Messick, 1989). He cogently follows to express that construct-irrelevant variances refer to variables unrelated to the construct being measured; it occurs when the test scores are influenced by factors irrelevant to the construct. For example, (Bachman, 1990), an individual’s background knowledge, personality, characteristics, test-taking strategies, and general intellectual or cognitive ability all can be construct-irrelevant and effort needs to be made to keep influences such as these to a minimum. Thus, variables that systematically (rather than randomly) interfere with the ability to meaningfully interpret scores or ratings represent CIV.

There are always some unrelated-to--construct parameters which contaminate the measurement and also interpretation and use of test score. Gipps (1994, p. 46-47) touches on test score pollution or Lake wobegon effect, both of which, as construct irrelevant variances, are to the point here. By definition, score pollution refers to the practice of teaching to the test in order to raise test score. Attentively considered, increase in test score is not related to the construct; so, pollution is construct-irrelevant test score variance. By the same token, According to Schouwstra (2000), construct- irrelevant variance represents systematic interference in the measurement data, often associated with the scores of some, but not all, examinees. For example, to Bachman (1995) some poorly designed item formats can make it more difficult for some students to give a correct answer ; as he holds, DIF or DTF(Mousavi, 2009) are themselves CIVs into the measurement process.

If some students have prior access to test items and other students do not have such access, this type of test insecurity CIV makes score interpretation difficult or impossible and seriously reduces the validity evidence for the assessment as the same happens in teaching to the test; A clear-cut example is that Preparation Course for University Entrance Exam leads the teachers to just teach the content (Jipps,1994), finally, the 'test-wise' examinees obtain scores which are invalidly high for them (Messick, 1989). Furthermore, the instructor uses actual test items for teaching, thus creating misleading or incorrect inferences about the meaning of scores. Likewise, other types of test irregularities, such as cheating and also cramming for test when the learners by chance get encountered with some content on the test, in my idea, can give rise to score pollution and interpretation, therefore, to inappropriate decision making.

More evidently, there lie some extraneous parameters which overshadow the use, interpretation and consequences of the tests as well as the decision made founded on them. Clearly, content domain considered it has been frequently observed that test content, for example some reading passages of IELTS, is unfamiliarly technical and specialized so that some learners have background knowledge about it and some others do not, all of which can ultimately give rise to under-performance or over-performance with some learners compared to others. In parallel with CV, as one major threat to validity, is Construct under-representation (CU); it refers to the under-sampling or biased sampling of the content domain by the assessment instrument (Schouwstra, 2000). In a word, CU occurs when the test fails to capture important aspects of the construct that it is intended to measure or when part of the construct is not present in the test so that CU leads to underperformance on the part of test takers.

To avoid construct under-representation, one caution should be exerted with respect to drawing a parallel between test situation and non-test situation, i.e., authenticity, with use of test tasks which are not advantageous or disadvantageous, i.e., DIF and test bias, to the test takers, given their cultural, ethnic, geographical and educational background (Bachman & Palmer, 1996; Bachman, 1995). Hence, test items or tasks should be triggered to eliciting solely knowledge-dependent response on the part of the test takers. According to Messick (1989), construct under-representation refers to the imperfectness of tests in accessing all features of the construct, hence leaving out some important features that should have been included.

Bachman(1995, P.117-119) presents test method facets believing that, in addition to the abilities we measure, they also affect performance on the test and are effective to the validation of language tests; his five categories of facets include: environment, rubric, input, response, connection between input and response. Clearly, test method facets can be a source of construct irrelevant variance if the five categories are not more carefully regarded either at the design or administration stage of the test; for example, when the physical setting of group of test takers compared to other group's setting, in which test is administered, is much noisier, the response validity of the test-takers will be threatened.

Furthermore, Henning(1987) states that there are a variety of sources threatening validity of the test, including invalid application of tests, i.e., a test valid for proficiency purpose cannot bet valid for achievement, inappropriate selection of content, i.e, items are occasionally biased on favor of some particular language elements to the exclusion of others do not match the objectives or content of instruction, imperfect cooperation of the examinees, i.e., lack of response validity, inappropriate referent(norming population), i.e, items which are suitable for one group may function differently with another group, poor criteria selection, i.e, low validity of criterion-related validity underestimates true validity, sample truncation, i.e, artificial restriction of the range of ability results in underestimation of validity and reliability, use of invalid tests, i.e, situation in which the construct is not theoretically and operationally well-defined and this leads to invalid tests.

It is also worth citing that since reduced reliability is threat to validity, therefore, threats to reliability can be threats to validity; this is an accepted notion, of which the required caution should be taken account. Bachman (1995) puts the perspective that a test cannot be valid without being reliable. Therefore, the following fluctuation factors threatening unreliability must be also taken into account more seriously: learners, scoring, administration, test characteristics, response characteristics such as guessing and being test-wise (Brown, 2004).

6. Methodology

6.1. Research Questions and hypotheses

This study directs aim at investigating the impact of construct-irrelevant variance on test performance which leads ultimately to performance under-representation within the framework of the following research questions:

- Is there any difference between the anxiety level of two groups of test takers (math more-interested-in test takers and math less-interested-in test ones) associated with GRE math section?
  H1= There is no difference between the anxiety level of the two groups associated with GRE verbal section.
- Is there any difference between the performance of two groups of test takers (math more-interested-in and math less-interested-in test takers) on GRE verbal section?
  H2= There is no difference between the performance of two groups of test takers on GRE verbal section.
- To triangulate the data collection process, one descriptive construct-eliciting prompt being conducive to attitude-triggered response on the part of the test takers was introduced. As such, the test task presented to both of the groups similarly appears below: *What is your idea about the usefulness of the verbal section of GRE? Did the math section make you anxious?*

6. 2. Participants

The present study consisted totally of 80 test takers divided into math more-interested-in-test takers (N=40) and math less-interested-in test takers (N=40), without any reference to gender difference consideration. They were selected from two branches of Iranian Institute, in Ardebil, Iran. The age of the participants varied from 18 to 26 and also they were from various academic educational backgrounds and with respect to their English background, they were at advanced level. Accordingly, in sampling stage, an attempt was made to select the test takers at advanced level, those covering 1100 words and Essential Vocabulary for GRE as was according to the present research objectives instructed to the learners. Following that, the main aim for this was to get assured of the entrance behavior of the test takers in as homogeneity terms as possible associated with their GR-required lexical competence and also these words, as homework, were tackled and asked in the class, however most of them, as is displayed in the self-reported response from test takers, did not like the GRE=loaded vocabulary.

Moreover, they were never informed of the fact that they will take GRE test so as to avoid the interference resulting from test-wiseness and test preparation which normally contaminate the interpretation of score meaning and application of test for decision making purposes.

By the same token, the test takers were not cognizant that their math section won't be considered in research; after they filled in the Anxiety Questionnaire, took GRE verbal test and finally, gave an open ended response to descriptive task, they were announced that the math section was not included; I presented the math section to just truly observe their response validity and control the variance threatening response validity; so, it was controlled as a construct-irrelevant variance and as a measurement error. Also, to control their test wiseness, I used the test takers who had not already taken GRE test and that they were not even familiar with test structure. This was intentionally done to control test wise related construct irrelevant variance.

6.3. Instruments

Three instruments including GRE Anxiety Questionnaire, GRE-based verbal proficiency test and a self-designed descriptive task were employed in the present study. GRE anxiety questionnaire included 30 questions and the test takers selected either one of the following: not at all, valued one; a little: valued two; a fair amount, valued three; much, valued four and very much, valued five (see Appendix 1). A point worth citing is that the total score for the 30 items of Anxiety Questionnaire was 150. So, the scores gained by the test takers, as is evidenced in Table 1, is considered out of 150. GRE-based verbal proficiency test included 40 which totaled at 20 marks, i.e., every prompt scored 0.5. Furthermore, the third instrument demanding a descriptive task was administered similarly to both groups, targeted at triangulation for assuring the validity of the quantitative data-relayed interpretation.

Concerning the reliability of both verbal section and self-designed anxiety questionnaire with a view to face validity, it is necessitated to be voiced that since the test tasks associated with verbal section were extracted from Barron's GRE as an already-standardized test, the reliability of verbal section was taken for granted. On the other hand, the reliability of self-designed questionnaire was indirectly supported with a view to open-ended task which was administered to all the test takers in both groups, carrying the sense that, as the findings indicate, approximately the same response was elicited from the test takers. So, the reliability check in the research was supported by a previously standardized test, Barron's GRE and using two different methods as regards the self-designed questionnaire, indicating that the estimates of response by the groups were virtually the same.

6.4. Procedure and Design

For accomplishing the intended afore-cited objectives, a descriptive type design was employed. To this global aim, first, the participants were categorized into two groups of math more-interested-in test takers and math less-interested-in test ones so that with use of a mini-classroom- based interview their preferences and interests were reported and finally noted down. Then, the two groups' attitudes with respect to anxiety associated with face validity and math section as well as the effectiveness and practicality of verbal section were quantified through a self-designed questionnaire. So as to get assured of the validity of the answer, i.e., so as to reach the certainty that the filled-in questionnaire, i.e, research question one, has been honestly dealt with and also in order to get assured of the reliability of the self-designed questionnaire, triangulated approach in the form of a descriptive task demanding an open-ended response was intended and employed for two main purposes. Next, the test takers' descriptive responses and questionnaire responses were analyzed with use of *t-test.* In the end, the research questions were tackled and the mentioned hypotheses were either rejected or supported.

6.5. Data analysis and findings

Table1: below is a statistical elaboration on the first research question:

Groups	N	Mean	Std. Deviation	t	df	Sig.(2-tailed)
A) Math less-interested-in Test takers	40	99.1250	13.00530	8.000	78	.000
B) Math more-interested-in Test takers	40	69.8500	19.14392		68.677	

Table 1: T-test of anxiety associated with face validity and math section

As Table 1 above displays, some statistically significant differences ($t = 8$; $p < 0.0001$; Mean= 99.1250 and 69.85) were found out in the anxiety of the two groups; in the sense that, group A had higher anxiety compared to that of group B. A point worth citing is that the scores gained by the test takers, evidenced in Table 1, is considered out of 150, but in reverse order: the higher score indicates a negative feeling, i.e, higher anxiety and lower scores suggest a positive mood, i.e, lower anxiety.

In support of this, a self-reported response to the descriptive prompt appearing in research question section is illuminating; thus, this finding together with the following explanation totally delineates the rejection of the first hypothesis. Hence, there is a statistically significant difference between the anxiety level of the two groups associated with GRE verbal section as also evidenced below

First of all, the explanatory response from Group A portrayed that 34 test takers, i.e., 85%, less or more wrote " I got anxious once I saw the math section and I think this will affect my score and that the words are supposed not to be practical and useful in social setting"; one of them, i.e., 2.5%, wrote that it does not matter; and five of them, i.e., 12.5 %, had virtually written that it made no difference to them because the words are without any use. Furthermore, the open-ended attitudes of Group B proved the following: 26 test takers, i.e., 65%, responded that "math section had a pleasant effect on me or so and they however stressed the uselessness of the words on GRE verbal section"; five of them, i.e., 12.5 %, wrote " it was OK or so and the words were strange"; two of them, i.e., 5%, had written "I have no idea" and finally, seven of them, i.e., 17.5 % wrote that it made no difference to them and they did as they could; however, these seven test takers had the same idea with respect to the uselessness and impracticality of GRE verb section. Viewed statistically and triangulated technically, the findings are supportive to each other. Evidently, the findings are compatible and supportive to each other, however not identical since both of them reject the first hypothesis.

Additionally, to supply answer to the second research question, Table 2 below is numerically more informative:

Table 2: T-test of verbal-section performance of the test takers

Groups on verbal section	N	Mean	Std. Deviation	t	Sig.(2-tailed)
A) Math less-interested-in Test takers	40	4.9000	3.52136	8.336	.000
B) Math more-interested-in Test takers	40	11.6750	3.74431		

As Table 2 reveals, the second research question reached a statistically significant difference (*t* =-8.336; *p* < 0.0001; Mean= 4.90 and 11.67); the drawn conclusion pursues that group B outperformed group A on GRE verbal section; thus, the finding leads to the rejection of the hypothesis. Therefore, there is a statistically significant difference between the performances of the two groups of test takers on GRE verbal section.

6.6. Discussion and conclusion

The results of the findings appear, the means and Sig.(2 tailed) noticed, in Tables 1 and 2 together with supportive self-reposted descriptive response. Remarkably, Table 1 indicates the statistically significant difference (*t* = 8; *p* < 0.0001; Mean= 99 and 69) between the two groups with respect to anxiety. The point more arguable is that, based on the response they have provided through open-ended prompt, they have considered the math section as construct-irrelevant variable as arises from face validity. So as to ensure of the response validity of the test takers in tackling the questionnaire on the one hand and to estimate the reliability of the test on the other hand, an explanatory response was also elicited. Indicatively, opposite fit was observed between the attitudes of the test takers, 85%, less or more wrote " I got anxious once I saw the math section and I think this will affect my score and that the words did not seem practical and useful in social setting". Along the same line with the strangeness and impracticality of the vocabulary test of GRE and in opposite line, expectedly, with the math section, group B test takers, 65%, responded that "math section had a pleasant effect on me or so and they however stressed the uselessness of the words on GRE verbal section and the others fell in between.

Vividly present and statistically supported, group B outperformed group A on GRE verbal section as a function of anxiety variable leading to measurement error. Therefore, performance under-representation was observed which raises a hot argument as to the fact that the subject-nature design of GRE can be more reliable and valid than the intelligence-based design of GRE. On the other hand, If we are intended to test the abilities of the learners in terms of their intelligence, why should we do it with use of vocabulary items which are of little practicality? Put another way, does GRE have any social consequence? Have the value implications, evidential foundation and ethical issues been considered on GRE?

If we are bound to measure the intelligence and analytical abilities of the test takers with respect to their lexical competence, for academic purposes, is it urgently required to do this through strange words, the math section taken for granted and accepted, for example? If scores on GRE indicate the abilities of the test takers, is not this contrary to the modern model by Gardner (1983) and Armstrong(2000), i.e., multiple-intelligences theory including more than seven intelligences and referring to the rejection of intelligence in singular form and intelligence test to the inclusion of intelligences in plural form? Has not intelligence test been rejected for being static and unit-faceted, instead of which multi-faceted and dynamic view of the intelligences (Gardner,1983; Armstrong, 2000) have been proposed and performed? Are not intelligences changeable according to contextual and cultural variables affecting it? If intelligence is not fixed, but changeable and of potential to grow up, why is GRE established on intelligence-based design rather than on intelligence**s**-based design (intelligence versus intelligence**s**)?

Thus, the future-research-generating implications inspired by the present study is tackled in a way as to either the intelligence (in singular form, i.e, traditional view of intelligence) of the GRE applicants is included in a single score indicating their ability or that their intelligences(plural form, i.e, modern view of intelligence) are considered and interpreted in context? How the usefulness and appropriateness of GRE-driven single score justify the value implications and social consequences? Clearly put, what is the main purpose of GRE vocabulary section? Are these words needed practically? Will the applicants be exposed to them in non-test situation? If its purpose is to measure the intelligence of the test takers, how is it justified with a view to Gardner's multiple intelligences theory?

In a word, I would hasten to point out that living with this contradiction is, however, blessed with paving the way for the future research. In other words, the answer to these questions is ultimately left to the researchers and testers in in the field, the academics who can contribute to settling down the argument associated with GRE verbal section and also the need to separate the verbal section from math section.

References

Armstrong. A. (2000).*Multiple intelligences in the classroom*(2nd ed). Association for supervision and curriculum development.

Bachman, L. F. (1995). *Fundamental Consideration in language testing.* Oxford: OUP.

Bachman, L. F. (1999). Language Testing at the Turn of the Century: Making Sure that What We Count Counts. *AAA Letter, 20*(2). Retrieved from the World Wide Web: http://aaal.lang.uiuc.edu/%20letter/testing.html

Bachman, L. F. & Palmer, A.S. (2000). *Language testing in practice.* Oxford: OUP.

Birjandi, P. & Mossallanejad, P. (2010). *An overview of testing and assessment: from theory to practice.* Sepahan Publication

Brown, H. D. (2004). *Language assessment: principles and classroom practices.* San Francisco State University

Canale, M. (1984). Testing in a communicative approach. In G.A. Jarvis (Ed.), *The challenge for excellence in foreign language education* (pp. 79-92). Middlebury, VT: The northeast conference organization.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.

Carroll, J. B. (1972). *Fundamental considerations in testing for English language proficiency of foreign students.* In H. B. Allen, & R. N. Campbell (Eds.), *Teaching English as a second language*: A book of readings (2nd Ed.). New York: Mc Graw-Hill.

Davies, A. (1990). *Principles of Language Testing*. Oxford: Basil Blackwell, Ltd.

Farhady,H.(1983a). *On the plausibility of unitary language proficiency factor*. In WJ.Oller.Jr.(Ed.), Issues in language testing research. Newbury House Publishers.

Farhady,H.(1983b). *New directions for ESL proficiency testing.*. In WJ. Oller.Jr.(Ed.), Issues in language testing research. Newbury House Publishers.

Fulcher,D. & Davidson, F. (2007). *Language Testing and Assessment*. Routledge Applied Linguistics.

Heaton, J. B. (1988). *Writing English Language Tests*. Longman Group UK, Limited.

Heaton, J. B. (1991). *Language Testing*. Hayes, Middx: Modern English publications

Henning, G. (1987). *A guide to language testing: Development, evaluation, research. Rowley, MA: Newbury House.*

Hinofotis, F. B. (1981). Perspectives on language testing: past, present and future. *Nagoya Cakuin, Kyoiku kiyo, 4,* 51-59. http://www2.hawaii.edu/~roever/wbt.htm

Hymes, D. H. (1967). Models of interaction of language and social setting. *Journal of Social Issues*, 33, 8-28.

Gardner,H.(1983). *Frames of Mind: Theory of multiple intelligences*. NewYork: Basic Books.

Gipps, C.V.(1994). *Beyond Testing: towards a theory of educational assessment*. Palmer Press.

Kelley T.L. (1927). *Interpretation of educational measurements*. Yonkers, NY, World Book Company.

Messick, S. (1996). Validity and washback in language testing. *Language Testing 13 (3): 241-256.*

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

Messick, S. (l989b). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Newton,P. & Shaw,S.(2013). *Book Announcement: validity in Educational and psychological assessment.*Research Notes, Issue 54, November 2013, Cambridge English Language Assessment.1913-2013.

Oller, J. W., Jr. (1979). *Language tests at school: A pragmatic approach.* London: Longman.

Oller, J. W., Jr. (Ed.), (1983). *Issues in language testing research*. Rowley, MA: Newbury House.

Oller, J. W., Jr. (Ed.), (1986). *Communication theory and testing: what and how.* In Stansfield (Ed.), (pp. 104-55).

Savignon, S. J. (1983). *Communicative competence: Theory and classroom practice*. Reading, MA: Addison-Wesley.

Schouwstra, S.J.(2000). *On Testing Plausible Threats to Contruct Validity).* Downloaded from UvA-DARE, the institutional repository of the University of Amsterdam (UvA). http://dare.uva.nl/document/56520

Spolsky, B. (1978). Introduction: Linguistics and language testers. In B. Spolsky (Ed.), *Advances in language testing series*: 2. Arlington, VA: Center for Applied Linguistics

Weir, C. J. (1990). *Communicative language testing.* Prentice-Hall, Inc.

Zhang,X.(2013). Foreign language listening anxiety and causal relationships. *System 41(2013)164-177.*