A Separate title page

Title: Conceptualizing Essay Tests` Reliability and Validity: From Research to Theory

Author Name: Badjadi Nour El Imane

Institution: Huazhong Normal University, China

Date of Creation: March 17th, 2013

The paper

Abstract

In the current paper on writing assessment surveys the literature on the reliability and validity of essay tests. The paper aims to examine the two concepts in relationship with essay testing as well as to provide a snapshot of the current understandings of the reliability and validity of essay tests as drawn in recent research studies. Bearing in mind that essay tests are the most crucially and widely used direct writing assessments worldwide, our research is driven by the fact that successful implementation of assessment practices depends on an understanding of reliability and validity. The review focuses primarily on the general theoretical and practical aspects of reliability and validity and particularly on the evolution within and the relationship between the two concepts. Aspiring to synthesize research findings about the validity and reliability of essay exams as a means of direct testing of writing, we focus on the three main axes raters, scoring scales, and the prompt of an essay test. The results would inform test developers and language testing researchers about the current status and future directions of writing assessment research. Research of this kind is needed so that a fuller understanding of the validity and reliability of essay exams is achieved.

Key words: essay testing, validity, reliability

1. Introduction

Direct writing assessment abides to be an inspiring and challenging venue for research in both first and second language. The issues discussed in this paper deal with the pressing concerns for impromptu essay testing: raters, scales, prompts, and their interaction with reliability and validity. The paper addresses the need for synthesis and review studies as the concepts of reliability and validity has evolved and ample research has been spearheaded, especially over the last two decades. Noticeably, recent research interests had extended from focusing mainly on the statistical aspects of writing assessment to more situated stances on contemporary writing assessment theory and practice. Moreover, assessing writing proficiency has always been deemed problematic for language educators and educational institutions. Though considerable amount of ink has been spilled on researching direct measures' reliability and validity, studies which focus jointly on both of these central concepts are few and far between. The paper sets with a brief introduction to essay testing as a strategy of direct writing assessment, its delineation, concerns, and chief goals. Then, an overview of reliability and validity as essential characteristics of efficient essay testing is presented. This is followed by a terse account of the three main theories akin to the psychometrics of educational assessment. We move then to the application of this latter to recent empirical investigations regarding essay test's development and rating. The paper ends with a snapshot of the prevalent landscape for essay testing reliability and validity as well as suggestions for further research.

1.1.Statement of the Problem

A reliable and valid assessment of the writing skill has been a longstanding issue in language testing. The nature the writing ability, the qualities of good writing and the ongoing need for writing in divergent fields all have whet the appetite for a better understanding of how this cognitively complex and linguistically multi-faceted skill can be measured. Thus, an influx of empirical research studies has been extensively conducted to gauge the reliability and validity of essay tests in relation with various variables and in varied contexts. However, though highly rigorous, their results seem fragmentary and are by nature partial. While synthesis studies are few and far between, they are necessary for capturing recent developments in the field from various angles. This paper is an attempt to provide an integrated vision of the reliability and validity of impromptu essay writing measures in relation with the three major themes associated with assessment procedure, namely, raters, scoring scales, and prompts. In addition, though the evolutions in perceiving reliability and validity is theoretically well established in the literature (e.g. Weir, 2005), few studies address the concordance between theory and research practice.

1.2.Research Questions

In pursuit for figuring out how recent scholarship in writing assessment portrays and employs the concepts of reliability and validity and the way they relate to each other. The paper is guided by the following research questions :

Why and in what ways is essay assessment' s reliability and validity measurement central to the improvement of language learning, instruction, and assessment?

What are the major considerations in writing assessment validation research?

What implications do the present-day conceptual status of reliability and validity?

2. Essay Tests as Direct Writing Assessment

Assessing candidates' language and writing skills is usually accomplished through direct or/and indirect measures. On one hand, indirect assessment uses objectively scored item types which focus on formal features usually at the word and sentence level. Indirect testing items target at assessing the formal aspect of writing such as grammar and vocabulary. This kind of items are viewed as more reliable due to their objectivity yet they are notoriously difficult to design and fall short of adequately assessing writing (Yancey, 1999).On the other hand, the direct approach requires candidates to compose written passage (s). Direct testing has higher validity since candidates are required to demonstrate their ability to write longer texts. Direct writing assessment procedures generally are based on the belief that validity is prior to though not independent from reliability. In sum, essay writing is a direct assessment procedure where examinees compose a piece of writing corresponding to prompts.

Moreover, direct tests are performance-based since they are designed to assess the testees' ability to perform a particular task. A measure is authentic when it reflects as closely as possible the construct it is designed to measure. Absolute authenticity is not absolutely attainable yet it can be improved through the use of direct tests. This latter are claimed to directly measure an ability by requiring performance akin to authentic language use (Davies et al. 1999). An essay test is a direct procedure for assessing candidates' writing ability. Essay tests are comprehensive tests which target at measuring knowledge of language as a whole not only knowledge of isolated language components.

Further, essay tests are used to assess candidates' ability write in a language. Writing is a cognitively complex task which encompasses a variety of sub-skills including grammar, spelling and vocabulary, discourse coherence, mechanics, clarity of ideas, content, and style (Ur, 1996; Weigle, 2002; Shrum & Glisan, 2000; Saunders & Scialfa 2003, Spandel, 2008; Wilson, 2006). One advantage which characterizes essay testing is its communicative orientation since it is not based upon the view that language competence can be measured through testing each of its constituting components separately. Besides, the major strength points of essay testing are, as stated by Crusan (2002), "that they are able to gauge the ability of students to identify and analyze problems, to identify audience and purpose, to argue, describe, and define, skills that are valued in composition classes." (p. 19). Further, since 1993, White viewed essay assessment and the holistic approaches to its scoring as the primary concern "When a university or college opens discussion of the measurement of writing ability these days" (p.89). These days as well, essay measurement remains at a central issue in writing assessment. Along with their usefulness for encompassing all the facets of the writing ability, impromptu essays are claimed to have highy validity and cost efficiency, especially if holistically scored, as well as relatively acceptable reliability under alert management of rating (O'Neill, 2003). Therefore, the use of essays as an assessment procedure has been a promising area of research into the reliability and validity of writing measures.

The same as other types of direct tests, essay exams are subject to subjective assessment and reliability problems. Differences among raters concerning which aspect of writing is more prior and their perceptions on what characterizes 'good' writing (O'Neil, Moore, & Huot 2009), along with the subjective nature of judging a piece writing (Davies et al. 1999), all further complicate the reliability and validity analysis of a measure. As a result, this latter has long been subject to debate and in unceasing need for validation (Williams, 1970; Williamson & Huot, 1992). Further, Backon's (2003) study demonstrate that both multiple choice and short response test items have similar convergent validity and reliability coefficients. However, using short responses to assess the writing ability rides roughshod over the discursive and rhetorical aspects of writing. Hence, researchers advocate that the use of both direct and indirect testing items to enhance construct validity (Messick 1993). Nevertheless, this solution does not solve writing test's low reliability and further complicates the analysis of a test's reliability. Besides, not only raters, but also prompts and rating can be sources of error (Huot,1990; O'N eill, 2003; Schoonen, 2005). Influences on essay exams' reliability and validity mainly draw from (1) the different opinions among raters about the characteristics of good writing (2) the accuracy and usefulness of holistic vs. analytical scoring procedures (3) the difference in quality of examinee's writing from one topic to another. Therefore, in this paper, it is held that reliability depends on the measure's readers, the way it is intended to be evaluated (scoring procedure), as well as the way it is constructed (topic and wording of prompts).

2.1. Essay Tests` Validity

Validity is a complex construct which has constantly triggered confusion and debate in the literature of educational measurement. Describing validity, Harrington (1998) states, “a valid assessment is one which assesses what is sets out to assess” (59). Similarly, Yancy (1999) explains, “validity means that you measure what you intend to measure”(487). Differently stated, according to Borrowman (1999), validity is about the connection between what a test claims to measure and what is actually measures. All of these definitions imply the relative nature of validity.i.e, that no test is absolutely valid. While there is no test which does not assess what it intends to assess; all the tests assess some of, most of, or other than, what it claims to be assessing. Likewise, a connection between what is claimed and what is actualized is similar to that between what is ideal and what is real. Further, according to Messick (1989), validity research uses “integrated evaluative judgment” supported by adequate and appropriate inferences based on test scores and modes of assessment (p.5). Broadly speaking, validity is the extent to which a test actually measures what it is designed to measure. Though there are numerous varieties of validity, the latter is usually delineated in light of four types of validity: content, construct, concurrent, and predictive; these latter are often jointly referred to as criterion-related validity. However, according to Messick (1994), since all of validity types, in one way or another, seek to provide evidence that a measure actually assesses the target trait or skill, they can be viewed as different aspects of construct validity.

First, internal validity is related to the content and characteristics of the test items and their corresponding responses. Content validity is a conceptual non-statistical validity centered on the analysis of a measure’s content to determine the extent to which it represents the knowledge or ability to be assessed. Content validity denotes how relevant and representative the test is in terms of covering the language ability and skills it intends to measure and it is increased through test specification. Construct validity refers to the extent to which a test echoes the theoretical doctrine it claims to be based on. It involves figuring out how test scores can be interpreted in relation to the theoretical framework underlying the construct a measure is intended to test. Construct validity involves gathering influence and making inferences in light of candidates’ performance. It can be measured through exploring the relations between an empirical findings and the theoretical explanatory concept of the construct under test (Davies et al,2002). Factor analysis and multi-trait multi-method analysis are often used to measure construct validity.

Second, criterion-related validity (also external validity) is related to the construct being tested or the criterion to which test performances are related. It is investigated through relating test scores to other such as teacher assessment (concurrent validity) or future achievement (predictive validity). Concurrent validity concerns the relation between a newly developed test and an already existing "criterion measure" such as a standardized test. If the two measures are related i.e. their results are similar, then the new test has concurrent validity. Predictive validity is centered on the extent to which a measure's results indicate performance on an external yet related criterion. Predictive validity is necessary since a test is likely to inform about performance in real world; it links language learning to language use.

Investigating validity may require correlational studies such as factor analysis and path analysis. For instance, Multi-trait multi-method analysis is an experimental design used for determining the extent to which scores can be linked to either candidate traits or to the effects of the testing method. Campbell and Fiske (1959) statistical method is related to construct validity. Correlations between different measures of the same ability should be highly positive (convergent) while the correlation between various traits measured should be low. Among the widely used methods are path analysis and factor analysis, yet what correlations to investigate depend on the focus of the validation study. Validity is prior to but reliant on reliability (Davies et al.1999).

The validity of a writing proficiency test is established by finding out the extent to which it accurately reflects the abstract concept of essay writing ability. Strictly defined, validity "refers to the inferences made about a test score, i.e., the degree to which it is useful as a measure of a particular trait for a particular purpose and for a particular examinee." When assessing validity, the test is analyzed to judge whether it addresses all the sub-skills or aspects of the target construct and whether it is measuring those sub-skills adequately. In essay exams, validity is usually addressed in terms of content, organization, mechanics, and language use. Nevertheless, the assessment of what is to be demonstrated in an essay response may vary across rating scales and what raters focus on in a particular field or writing genre. Huot (1996) suggests that the validation of a writing measure should include backing on theoretical foundations related to writing instruction as well as empirical data of writers' performance. He further argues that an investigation of a writing measure's validity and reliability should go beyond the statistical results to draw on and contribute to knowledge on how a construct should be taught and assessed.

2.2. Reliability

Unlike validity which is perceived as complex and thus less frequently examined, writing assessment reliability has always received enthralling interest (Huot & O' Neill, 2006, Williamson, 1993; Huot, 2002; O'Neill, Moore, & Huot, 2009). Reliability is essential yet not sufficient to establish a measure's validity.

To start with, reliability concerns the agreement between the results of one test and the test itself or another test. Reliability is defined in the Standards (AERA, APA, & NMCE,1999) "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be repeatable for an individual test taker"(p.180). Moreover, reliability encompasses the" degree to which scores are free of errors of measurement for a given group"(p.180 qtd in O'Neill, 2011, p 12 of 19). on the opposite, a measure is judged as unreliable when various measurements fail to provide consistent results. Reliability is thus referred to as the "steady-state requirement". Reliability is usually attributed to the measure itself or to the way it has been scored and, to a less extent, to the testing conditions. While the factors related item selection are most important for objective items tests, rating bias is particularly relevant to the reliability of subjective testing items. Hence, examiner bias is the most prevailing type of reliability akin to essay testing and is referred to as inter-rater reliability. However, the factors related to the testing conditions are rarely used for measuring reliability because of the difficulty of having the same candidates to sit the same exam and the effect of learning resulting from taking the test previously. Suggesting that even well-established measures need to be proved reliable for every new use, Buck (1992) contends: "There is no such thing as a test method which automatically produces reliable and valid tests, nor is there ever likely to be one. Each new test, or each new use of an old test, needs to be validated anew, and that naturally includes estimation of reliability" (p. 141). Moreover, reliability analysis is defined and framed mostly in terms of technical statistical operations. This understanding has drawn on the prevalent interest in objectivity in psychometrics during the early and mid-twentieth century (Williamson,1993, 1994).

Research echoes the recent calls for freeing validation research from the "reliability ritual" (Moss, 1992; 1994; 1995; Schils et al., 1991) and contextualizing reliability (Huot, 1996; O'Neill, 2011). The current epistemology of reliability tends to consider it as one aspect of validity or as "scoring validity" (Weir, 2005, p.1). Rather than rejecting the traditional perception of reliability, through "reframing reliability" (Adher-Kassner & O'Neill, 2010), contemporary trends tend to stretch the usefulness of reliability coefficients, standard error measurement, and similar statistics into accounting for broader field-specific, social, and critical issues. Furthermore, clarifying the relationship between reliability and validity which O' Neill (2011) suggests that validity is associated with the nature of the construct a measure intends to assess while reliability is focused the measure's potential to yield consistent results across various replications. Nevertheless, the relation between the two concepts remains a subject of confusion and debate among educational measurement experts and researchers, yet the rigorous states psychometrics occupy in assessment validation continues to be central.

2.3. The Psychometrics of Language Assessment

The evaluation of reliability and validity is psychometrically measured mainly using three theories: classical test theory, generalizability theory, and item response theory.

First, in Classical Test Theory, the evaluation of reliability involves test-retest, alternative forms, and internal consistency. Test-retest method draws on the consistency of scores when administering the same measure to the same group of candidates. The main drawback of test-retest procedure is that it assumes that no change (learning/forgetting) in testees' knowledge takes place if the measure is to be reliable. In addition, alternate forms (parallel forms) reliability involves developing two or more measures having the same specifications in terms of language and skills. Likewise, internal consistency requires that parallel items are constructed which is not compatible with essay tests since writing essays is time consuming and it is hardly possible to have the testees to write two essays without being destructed by psychological factors such as memory, concentration, and tiredness. Given that essay testing is an assessment which values the communicative value of texts as well as awareness of audience, the use of internal consistency for establishing reliability is argued to be problematic (Swain, 1993). Besides, it is scarcely possible to be sure two essays are truly analogous.

Second, according to Generalizability Theory (Cronback et al. 1972), the interpretation of generalizability coefficients also draws on the steadiness of scores and expects a specific level of consistency. It can be considered as the result of joining Classical Test Theory (CTT) and Analysis of Variance (ANOVA) and as a remedy for the limited efficiency of CTT inability to detect sources of variance. In addition, it is claimed to account more thoroughly for the factors influencing reliability than accounted for using CTT. GT further allows for detecting both the source of error and its effect on the consistency of scores (Davies et al.,1999). Nevertheless, GT theory is criticized for not being to yield sufficient mechanisms to back up judgments about interpretation (Nichols and Smith, 1998). The solution for low reliability according to this theory is a larger number of items (increasing test length) or more raters.

Last but not least, Item response theory also relies on the consistency of responses to achieve acceptable fit. It is used to estimate a candidate's ability through generalizing from a measure's results as displayed in performance on the writing task, along with item criteria and testee's traits included in the generation of data. The strength point of IRT is that it enables obtaining stable accounts of examinee's ability levels. This advantage permits further operations to be used such as test equation and Computer Adaptive Testing (Davies et al. 1999). Item response theory (IRT) is reflected in test construction and analytical methods based on the assumption of the uni-dimensionality of the construct being assessed (Henning, 1992). However, multi-dimensional models of IRT have been developed (Ackerman, 1994). According to IRT, reliability can be increased through standardizing testing conditions, lengthening the test, and using better testing items (Davies et al. 1999).

However a number of measurement experts and researchers doubt the reliability of a reliability analysis which does not account for the rationale of the learning theories in a particular domain (Linn et al. 1991; Messick 1994, Nichols & Smith, 1998). Instead, they argue that the measurement of reliability depends on the assumptions about learning and performance in a specific knowledge area or skill. Their view toward reliability seems to blur the distinction between validity and reliability. They further dismiss the notion that a reliable measurement may or may not be valid while an unreliable test may never be valid. Differently stated, " … a test may not have meaningful reliability without validity" (32). Nichols and Smith (1998) further criticize drawing solely on a test theory to measure consistency in performance. Instead, they suggest that reliability should be measured according to the theories of learning and characteristics of performance in a particular domain because neither of CTT, GT, or IRT provide a basis for expecting consistency or inconsistency across variant conditions and subjects. Rather, what they claim is that it is only after a domain is delineated that consistency or inconsistency can be attributed to deviance or error. Similarly, O'Neill (2011) argues:

> Writing is a complex, multidimensional, and contextually situated activity. Importing psychometric theory and practices, especially in terms of reliability, may undermine the very usefulness of a writing assessment's results. However, psychometric theory cannot be dismissed out of hand; instead, writing assessment scholars and practitioners need to draw on language, literacy and psychometric theories as well as other interpretive traditions to design assessments (p 9 of 19) .

Therefore, reliability can be measured only with reference to the construct a test measures or to the essential principles underlying the assessment of this construct.

3. Validation Research

Validity inquiry needs to focus on the purpose and use of the test's results and requires more than a quantitative analysis of the results. Indicating the importance of reliability and validity measurements on the part of teachers, Weigle (2007) states "Teachers should not hesitate to ask questions about the reliability and validity of the tests that their students are required to take and how test results will be used, and teachers should be proactive in bringing issues of questionable testing practices to the attention of administrators." The process of test validation involves collecting information about the validity and reliability of a test it terms of its fulfillment of the purpose it is designed for and scores' consistency on the basis of evidence derived from the scores. Besides, "Validity research involves a dynamic process that requires an examination of procedures and results, use of this information to revise and improve assessment practices, and an examination of revised practices in a never-ending feedback loop."

(O'Neill; 2003, 51). In other words, the validation of a measure takes into consideration the linguistic skills, the conceptual content, and candidates' test responses. There is no 'one way' to measure reliability and validity; studies use a proliferation of statistical and analyses mainly within three psychometric realms.

3.1.Research on Raters and Rating

For many years, writing assessment researchers have been interested in the reliability of rating. Certainly, inter-rater reliability and scale efficiency are the most examined aspect of writing assessment validation (Barkaoui, 2008). The effect of rater and rating has been heavily researched recently (Huot 1990; Schoonen et al 1997; Nichols & Smith, 1998; Weigle, 1994, 1999; Carr, 2000; Shi 2001,Knoch, 2009; Bacha, 2001; Lumley 2002; Sakyi, 2000; Schaefer, 2008; Johnson & Lim 2009, Barkaoui, 2010, Cumming et al. 2002) all of these studies seek to uncover the processes underlying the rating practice and emphasize the importance of rater training and the use of appropriate scales. Other factors include the differences among raters application of scoring criteria as well as their linguistic backgrounds and rater experience.

Two-fold studies are also used in investigating the reliability and validity of essay exams. For instance, Weigle (1999) examined variance in essay scores yielded by experienced and inexperienced raters across two different prompts. She found that differences were akin to the ease with which the two groups of raters employed the scoring rubrics with the two prompts. Her findings highlight the high potential and fruitful results of integrating quantitative and qualitative procedures for assessment validation.

In addition, Gamaroff (2000) suggests that inter-rater reliability is most important factor influencing an essay testing measure. He views that it is mainly influenced by the priority given to by raters to different aspects of writing along with the agreement about what should be considered prior. He further suggests, "Validity and reliability are two sides of the same corner" (p.44). Gamaroff thus argues that a rater's background does not hinder a valid and reliable judgment of an essay response since non-native raters are not necessarily less professional assessors.

Further, research demonstrated that rater-training can help, in achieving common stances, interpretations, and agreement (Weigle, 1994; Sakyi, 2000; Schoonen, 2005;

DeCarlo, 2005). Lumley (2002), in his investigation of four raters' application of a rating scale, has indicated that the rater decision-making involves a complexity the cognitive activities and affective factors. Research of similar concerns discusses the decision-making behaviors employed by experienced and unexperienced raters of candidates' essays. The methods used in rater-related studies included asking raters to clarify the criteria they perceive to be essential to effective writing and think-aloud protocols to discover the strategies raters use when evaluating an essay response.

Barkaoui (2007) used a mixed-method approach to investigate the effects of two different rating scales on EFL essay scores, rating processes, and raters' perceptions. G-theory was used to score the essays marked by four teachers with whom think-aloud protocols were used. Each rater used a holistic scale to score two essays silently and two others thinking aloud; then doing the same tasks using a multiple-trait rating scale. The result indicated that the holistic scale resulted in higher inter-rater agreement and that raters employed similar processes with both rating scales. Raters, not scales, were found to be the main source of variability. Opposite to what is widely held, the findings suggest that the holistic scale resulted in higher score reliability. The multiple-trait scale, on the other hand, resulted in lower score reliability because of the lack of training and thus can be viewed as having limited practicality in the study's context.

3.2.Research on Scoring Scales

The scoring criteria are deemed important in determining a test's reliability and validity. As far as composition tests are concerned, unlike the factor of raters which is mainly associated with reliability, or the factor of prompts which is chiefly linked to validity, the scoring scale links the two factors and thereby affects validity and reliability. Moreover, the scoring criteria are often referred to in terms of their purposefulness and accuracy in measuring examinees' writing ability. They are likely to reflect the test's purpose (s) and denote the aspects of writing it is purported to measure as well as to signify the characteristics of the test item, i.e. essay composition.

First, holistic and analytical approaches to scoring can be used to assess an essay. Multiple scoring by different raters may increase an essay measure's reliability since different scorers may chose to focus on distinct aspects of writing (Davies et al. 1999). In a typical analytic scoring method, a separate score is attributed to each feature or aspect of a composition task: content, organization and structure (vocabulary and grammar. This approach allows for diagnostic reporting of testees' literacy development, as well as to increase the test's reliability since raters are required to focus on the same aspects of performance. Besides, the same multiple-trait scoring procedure may be used for a variety of writing prompts which have similar test specifications. Test specifications refer to the documentation of what a test is intended to measure and how it is to do so as. They are important for achieving high construct validity because they include the test purpose, content, and format as well as the target population, the language of rubric, time span, and the scoring method.

In addition, providing separate scores for each aspect of the writing skill is an alternate for obtaining multiple holistic scores by different raters on a direct writing test item. On one hand, multiple-trait multi-method yields a more valid means for evaluation because it reveals testees' responses of various aspects and takes into consideration the fact that students are better at some aspects of writing than others. It is argued that giving a single global score obscures variation in performance within the writing skill. On the other hand, it is possible to combine holistic and multi-trait multi method approaches to scoring the responses of essay tests. Raters are asked to read twice: once, to holistically evaluate content and organization and then to analytically evaluate vocabulary, grammar, and mechanics. The overall judgment of a paper is calculated through totalizing single traits' scores. Nevertheless, this sort of multi-tasking the evaluation of an essay has low practicality particularly in case of large scale exams where the number of testees is usually huge.

Moreover, whether integer numbers or decimal numbers are used with a scoring scale is argued to affect inter-rater reliability. For assessments where various correct responses are accepted, Penny et al. (2000) suggest that raters augment integer-level scores by adding a fraction in order to improve inter-rater reliability in the scoring of performance. The decimal can be added when a writer's response is inferior or superior to the benchmark attributed to a given aspect of writing. Their study yields evidence that the augmentation integer scores increases inter-rater reliability. The rational they provide is that "true proficiency lies on a continuum that underlies the rating scale". Penny et al. 's study gives insights to the developmental nature of writing. Nevertheless, assessing inter-rater reliability in light of augmentation alone provides scant information when a deeper understanding of the scoring process and writing performance is required. Besides, suggesting that increased likelihood of variance error may be attributable to the high number of levels constituting a scoring scale, Penny et al. advocate that agreement among raters is more difficult to achieve for a five-point analytic scale than for a four-point holistic scale. Hence, the length of a scale seems to influence measurement error, yet whether the scale length may make the task of discriminating between levels of writing proficiency more demanding remains subject for further research.

Moreover, the holistic scoring of students' essays is widely used for its practicality. Huot (1990) found that inter-rater reliability vary across analytic, primary trait, and holistic scoring methods, as well as that holistic scales, though relatively have lower in inter-rater reliability, are more economical. A number of studies investigated the complexities involved in holistic writing assessment. Carell's (1995) work addresses the relationship between holistic scoring and raters' personalities. His research sheds light on the processes used by different types of reader-raters and their effect on the holistic scoring of writers' essays. Her research demonstrates that rater's personality type affects the score they assign. Besides, her study reveals a loose connection between the match of raters' personalities with writers' writing styles and the scores attributed to students' essays. The reliability of the holistic scale used in the study was analyzed using inter-rater correlation coefficients. The latter were used instrumentally as evidence or means for understanding the interaction between raters' personalities a scoring scale's reliability. Through highlighting the potential influence raters' personalities have on the valid application of a rating scale, Carell proposed that the reliability and validity of writing measures can be enhanced through raising raters' awareness to the influence their personality type may have of the evaluation of writers' performance.

Bacha (2001) compared the difference in the rating of essay responses using holistic and analytical scoring. The same rating instrument (Jacobs et al. scoring scale, 1981) was applied in two ways, holistically and analytically. Her study targeted to estimate inter-rater reliability through marking essay responses in different ways on various occasions. Bacha thereby concluded that an analytical scale is more informing for developing an appropriate EFL writing program. Though the focus of her research is not to indicate in what ways a scoring scale influences the reliability and validity of an essay exam, her study highlights the significance of using not necessarily distinct scales, but the same scale in different ways depending on the purpose of the measure and the validation process.

In a practical approach towards addressing the debate about the different methods used to assess reliability of essay testing. Sudweeksa et al. (2005) compared Generalizability Theory (GT) and Many Facet Rasch Measurement (MFRM). The aim of their study was to improve the rating scales used to evaluate testees' essays. They suggest that the two methods can be used complementarily since both have strengths and weaknesses and conclude that GT and MFRM should be used appropriately to the focus of a validation investigation.

Reinheimer's (2007) placement writing measure validation study reflects the relatively recent understanding of validity as an argumentative act. Though moderately positive, the study's results provide a framework for a validation process which separates but involves both scoring and validation (the study included scoring sessions followed by validation sessions) .after all the essays are holistically marked according to a 6-point scale, a sample from the essays is used for validation. This latter involves developing a scale from the program objectives, the principles of sound assessment, the writer's essays. His approach employs the already existing principles of how assessment 'ought to be' and local methods in order to improve program performance through the validity argument. The study sounds useful for developing an effective program review and provides evidence that perceiving validation as a rhetorical rather than a mathematical process can help improve writing programs and future assessments.

Similarly, East (2009) in an exploratory study focusing on interaction between scoring scale and a writing measure's reliability investigated the reliability of a rubric used in writing tests in two different test conditions, namely with and without dictionary. His research provides insights into ways of determining the reliability of scoring scales to be used in contexts where no more than two raters are available.

Exploring the effect the using order of holistic and analytic scales on reliability, Singer and LeMahieu (2011) investigated the independence of scores provided and processes used by readers. Raters were asked to evaluate the same set of papers in three ways: (1) holistically then analytically, (2) analytically then holistically, (3) holistically only or analytically only. The researchers indicate that when holistic scoring follows analytic scoring, the mean scores were significantly higher than the pure scoring. On the opposite, when holistic scoring was followed by analytic, mean scores were similar to the pure scoring. The research generally ascertains the assumption that raters score more validly when they first evaluate candidates' writing as whole then move to assessing its traits.

3.3.Research on Prompts

Tasks` instructions as reflected in the measure's prompts are considered to have a direct effect on a test's validity. Highlighting the significance of the effect prompts have on a test's reliability, O'Neill (2011) states: "if students' performances are not accurate in terms of their writing abilities because of the prompt design, then results are not reliable no matter how consistently raters apply the rubric and how much they agree with each other"(p.4 of 19).

One way in which prompts are claimed to influence a measure's validity is generality vs. specification. A specific topic prompts are useful in Language for Specific Purpose (LSP) writing tests. The purpose of LSP testing is not test content knowledge, though includes it, but to test language knowledge and the ability to write in a particular domain. As a result, the testees' population is likely to have homogenous background and similar language needs. Besides engaging the testees in capitalizing on both their knowledge of language and content, LSP tests relatively solve the problem of sampling in language testing since "It offers the prospect of exact specification of language features which make up a particular domain" (Davies et al., 1999: 104). As a result, they have high predictive validity because they address aspects of context and language use which influence performance. However, whether a domain related topic or a general topic influence LSP testees' writing performance on essay exams has not yet been adequately investigated.

From a different angle, Polio and Glew (1996) investigated writers' selection of essay topics according to the prompts they are presented with. They argue for providing options to writers and stating: “Denying students a choice may increase reliability, but it is possible that forcing them to write on a particular topic renders the test less valid”. Though seemingly indecisive their study indicates that few writers waste time by deciding to change the topic and that they are rational in their choices as they select essay topics which allows them to better display their writing ability.

In addition, studies investigating the application of the writing process to essay writing measures are relatively few. Lee (2006), for example, explored the difference in writers’ essays between first and second drafts composed while having opportunities for feedback, reflection, and revision during an ESL writing assessment. The final drafts were scored using analytical and holistic rubrics and inter-rater reliability coefficients for all aspects of the text analysis were above 0.9. Nevertheless, thought Lee’s study has no intention for the validation of this type of assessment, this latter is best, if not only, valid as formative assessment procedure. Similarly, Cho (2003) used a workshop-based essay test where writers are allowed to revise and receive feedback from other examinees. Given that the writing process is frequently associated with portfolio assessment which has a long way ahead to replace the widely used essay testing, the validation of process-based essay testing, though ‘promising’ may be challenging since students’ performance is likely to be different without the opulence of the opportunities and resources provided.

The use of graphs as prompts has recently received considerable interest among researchers investigating the relationship between task type and performance. In a validation study, Yang (2012) investigated graph writing strategies used by L2 learners. Writers’ essays were scored by two raters using analytical rubrics and the ratings were averaged using a third rater’s mark to increase inter-rater reliability. His study provides evidence for the substantive validity of the graph writing test task. He further found that most writers encountered lexis-related rather than syntax-related difficulties. Therefore, arguing for more valid evaluation of writers’ ability levels and more accurate detection of their areas of weakness, he suggests that lexical and grammatical abilities are likely to be regarded as two dimensions in scoring rubrics and thus need to be evaluated separately.

He and Shi (2012) found that the prompts used in direct writing tests significantly influence testees' performance on the writing task; and similarly topics (Lee & Anderson, 2007), as well as the integration of writing and reading (Gebril, 2009). Likewise, Lim (2010) explored the effects of prompts on a test's validity.

4. Conclusions and Implications for Further Research

The purpose of this paper was to present a practical research-driven rather than theory-driven conceptualization of validity and reliability. Three aspects are deemed central to the validity and reliability of essay exams: raters, tasks, and scoring methods. It is through a better understanding of the factors and criteria influencing this testing item's reliability and validity that a more efficient and valid essay exams can be constructed and appraised. The review reveals a new vision toward the two concepts, a vision which counts on usefulness and contextualization of writing assessment. It further uncovers a shift from focusing on deviance to an interest in variability (reliability) as well as from emphasizing accuracy or honesty to a recognition of relativity (validity). The distinction between reliability and validity persists in validation research for its practicality though the argument for considering reliability as one aspect of validity remains convincible and widely appreciated.

First and foremost, the traditional view toward reliability and validity as purely distinct has been challenged for failing to wholly capture the multi-faceted nature of language testing. Prompts of essay testing for instance are difficult to absolutely judge affecting a measure's reliability rather than validity. As O'Neill (2011) suggests, reporting Cherry and Meyer(1993), that "[identifying]differences in results across topics as reliability issue when in fact these differences are about validity. Variation across topics/prompts….can be a validity issue because the underlying construct being tapped is different if the writing tasks are different" (p 4 of 19). Therefore, there have been arguments for considering reliability as part of validity (Messick 1989; Nichols & Smith,1998; Broad,1994; Hout, 2002; O'Neill, 2003, Weir, 2005). Reliability and validity are interwoven and concerned with both of empirical statistical and theoretical conceptual aspects of the assessment process.

Consequently, decisions concerning prompts, which scoring scale to adopt, and how to use it depend on the purpose of the essay exam. Thus, the relation between reliability and validity is more intricate than to be accounted for in views of 'without' i.e.,' validity without reliability' or 'reliability without validity'. In addition, raters' development and adaptation of rating scales for their own assessment contexts largely guarantees that they use them properly and thereby provide more steady scores. Issues related to raters are explored in relationship with the scoring scale, the writing task, and the context in which the measure is administrated and the purpose for which a scale is employed. The shift in research on reliability and validity seems to be flowing from going beyond the clear-cut model of the two concepts and moving toward 'meaningful reliability'.

The concept of reliability has evolved into a hybrid construct constituting of both statistical psychometric and educational field-specific traits (see for example, Huot, 1996, 2002, O' Neill, 2013). Reliability includes though not limited to statistications of scores' consistency. The recent conceptualization entails an inter-related productive stance toward assessment validation research. This is embodied in the studies' efforts to use reliability coefficients for the purposes of the validation investigations, for understanding rating as reading, and for developing more valid prompts and rating scales. In other words, the question to be answered is no more whether or not scores are consistent, but what does consistency or inconsistency tell about the developent of a construct's assessment. Assessing a measure's reliability is not viewed as an end as one step toward achieving validation which is in turn leads to further ends concerning an understanding of writing proficiency, as well as the enhancement of its teaching and assessment.

Surprisingly, there are few attempts to understand writers and writing in writing assessment validation research. Though there is some research dealing with the effect of the way writers respond to a prompt on its validity, little is known about how writers come to compose the text to be read and rated; for instance, if writing is a process, what steps writers opt to focus on? Or what writers think 'good' writing is? And in turn what are the effect of their options and perceptions on their scores as well as the measure's reliability and validity? And what washback can be extracted to inform instruction in small-scale assessment and preparation for large-scale measurements? Moreover, the performance displayed by writers according to a particular type of writing or writing task is of enthralling interest in direct writing assessment research. Nevertheless, studies exploring the processes writers engage in response to different prompts are relatively scarce. Writing based on picture prompts is proved to influence performance (Yu, Rea-Dickins and Kiely, 2007). Is there a difference between visual, and different learning styles on performance; how reliable is this way of testing in comparison to purely linguistic prompts? How about the validity of open essay exams vs. controlled as demonstrated in writers' performance?

Further, there is a growing tendency in writing assessment validation research toward viewing writing performance as multi-faceted as shown in the multi-method research studies investigations of essay tests. One result of this tendency is that inter-rater reliability is being investigated in relationship with the scoring scale, the context in which the measure is administrated, and the purpose for which a scale is employed. The tendency further might have fruitful implications for SL writing research. With the development of new insights into SLA, tasks, content, cooperation, collaboration, and project work are argued to enhance the writing skill. "Current research in testing argues for a more direct connection between teaching and testing. The same kinds of activities can serve as valid testing formats with instruction and evaluation more closely integrated." (Shrum & Glisan, 2003: 292).

With the emerging trend in SL instruction towards process-oriented teaching of writing; the question regarding how to link instruction to testing is worth of further investigation, especially that currently used essay tests focus on the product while it is claimed that a measure aims at testing what was taught in the way it was taught (Shrum & Glisan, 2003). Last but not least, needs analysis is conducted can be integrated as a stage of the test development process to raise content validity since it provides a rational for selecting content. We thereby would like to highlight the importance of furnishing additional research on needs analysis-based writing tests.

Eventually, the paradigm shift in understanding and investigating reliability and validity reflected in the even attempts to 'free' reliability and validity from adherence to the 'reliability ritual' of psychometric bounds. Rather, these latter are viewed as means to an end not the ultimate end of validation research. Besides, variance is being seen not in terms of deviance hindering reliability but as a reality to be dealt with. This (r)evolution is theoretically described in Petruzzi (2008)'s philosophically rooted account of hermeneutic theory application to writing assessment. The paradigm shift in reliability research is portrayed in several studies' attempts to ‘understand’ the sources of inter（un）reliability. We thereby argue that the new conceptualization of reliability and validity is that of variability and relativity with focus on usefulness, contextualization, and situatedness of writing assessment practices.

Summing up, research echoes the recent calls for freeing validation research from the “reliability ritual” (Moss, 1992; 1994; 1995; Schils et al., 1991) and contextualizing reliability (Huot, 1996; O’Neill, 2011). The current epistemology of reliability tends to consider it as one aspect of validity or as “scoring validity” (Weir, 2005, p.1). Rather than rejecting the traditional perception of reliability, through “reframing reliability” (Adher-Kassner & O’Neill, 2010), contemporary trends tend to stretch the usefulness of reliability coefficients, standard error measurement, and similar statistics into accounting for broader field-specific, social, and critical issues. Furthermore, clarifying the relationship between reliability and validity which O’ Neill (2011) suggests that validity is associated with the nature of the construct a measure intends to assess while reliability is focused the measure’s potential to yield consistent results across various replications. Nevertheless, the relation between the two concepts remains a subject of confusion and debate among educational measurement experts and researchers.

References:

Ackerman, T.A., (1994). Using multidimentional item response theory to understand what items and tests are measuring. applied measurement in education. 7 (4), 255-278.

Bacha. N. (2001). Writing evaluation: what can analytic versus holistic essay scoring tell us? System 29 (2001) 371–383.

Bacon,D.R.(2003). Assessing learning outcomes: a comparison of multiple-choice and short-answer questions in a marketing context. Journal of Marketing Education, 25(1),31-37.

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

Barkaoui, K. (2008).*Participants*, texts, and processes in esl/efl essay tests: a narrative review of the literature.The Canadian Modern Language Review, 64(1), 99-134.

Barkaoui, K. (2010). Variability in ESL essay rating process: the role of the scale and rater experience. Language Assassment Quarterly, 7(1), 54-74

Borrowman,S.(1999).Thetrinity of portfolio placement: validity, reliability, and curriculum reform. Journal of Writing Program Administration,1(2),7-27.

Buck. G. (1992). Translation as a language testing procedure: does it work? Language Testing, 9 (Dec 1,). 123-148.

Carr,N.(2000).A comparison of the effects of analytic and holistic composition in the context of composition tests. Issues in Applied Linguistics,11(2),207–241.

Carrell, P.L. (1995). The effect of writers’personalities and raters’personalities on the holistic evaluation of writing. Assessing Writing, 2(21), 153-l 90,

Cho, Y. T. (2003). Assessing writing: Are we bound byonly one method? Assessing Writing, 8(2003), 165–191

Crusan, D. (2002). An assessment of ESL writing placement assessment. Assessing Writing 8(2003), 17–30

Cumming,A., Kantor,R., & Powers,D.E.(2002).Decision-making while rating ESL/EFL writing tasks:A descriptive framework.Modern Language Journal,86(1),67–96.

Davies, A., Brown, Elder, A. C., Hill, K., Lumley., T., & McNamara. T. (1999). *Dictionary of Language Testing*. Cambridge, UK: Cambridge University Press.

DeCarlo,L.T.(2005). A model of rater behavior in essay grading based on signal detection theory.Journal of Educational Measurement,42,53–76.

East. M. (2009). Evaluating the reliability of a detailed analytic scoring rubric for foreign language writing. Assessing Writing 14, 88–115.

Gamaroff, R.(2000). Rater reliability in language assessment: the bug of all bears. System 28,31-53

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

He. L., & Shi. L. (2012). Topical knowledge and ESL writing. Language Testing, 29(3), 443-464

Henning, G. (1992). Dimensionality and construct validity of language tests. Language Testing, 9(1), 1-11.

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

Huot,B. (1990).The literature of direct writing assessment: major concerns and prevailing trends. Review of Educational Research,60(2),237-263.

Huot,B. (1996).Toward a new theory of writing assessment.College Composition and Communication, 47(1996),549-566.

Huot,B.(2002).(Re)Articulating Writing Assessment for Teaching and Learning. Longman: Utah State University Press.

Johnson. J.S. & Lim. G. S. (2009). The influence of rater language background on writing performance assessment. Language Testing, 26(4) :485-505

Knoch. U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. Language Testing 26(2), 275-304.

Lee, H.K., & Anderson, C. (2007). Validity and topic generality of a writing performance test. Language Testing, 24(3), 307-330.

Lee, Y.J. (2006). The process-oriented ESL writing assessment: Promises and challenges. Journal of Second Language Writing 15: 307–330.

Lim. G. S.(2010). Investigating prompt effects in writing performance assessment. Spaan Fellow Working Papers in Second or Foreign Language Assessment, Volume 8:95–116

Linn,R.L., Baker,E.L., & Dunbar,S.B. (1991).Complex,performance-based assessment:Expectations and validation criteria.Educational Researcher,20(8),5-21.

Lumley, T. (2002). Assessment criteria in a large-scale writing test:what do they really mean to the raters? Language Testing, (3), 19:246-276.

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

Messick,S.(1993).Validity. In R.L.Linn(Ed.), Educational Measurement. 3rd ed. American Council of Education, New York. 13-103

Messick,S.(1994).The interplay of evidence and consequences in the validation of performance Assessments. Educational Researcher,23(2),13-23.

Moss,P.A. (1994). Can there be validity without reliability?Educational Researcher,23(2),5-12.

Moss,P.A.(1992).Shifting conceptions of validity in educational measurement:Implications for performance assessment.Review of Educational Research,62,229-58.

Moss,P.A.(1995).Themes and variations in validity theory.Educational Measurement: Issues and Practices,14(2),5-13.

Nichols. P. D., & Smith. P. L.(1998). Contextualizing the Interpretation of Reliability Data. EducationalMeasurement:Issues and Practice, 17(3), 24–36

O'Neill, P. (2003). Moving Beyond Holistic Scoring Through Validity Inquiry. Journal of Writing Assessment Vol.1,No.1,pp.47-65

O'Neill, P. (2011). Reframing Reliability for Writing Assessment. The Journal of Writing Assassment. 4(1), December 2011.

O'Neill,P.,Moore,C.,&Huot,B.(2009).A guide to college writing assessment.Logan,UT:Utah State University Press.

Penny, J., & Johnson.R.L., & Gordon, B. (2000). The effect of rating augmentation on inter-rater reliability: an empirical study of a holistic rubric. Assessing Writing 7, 143-164.

Petruzzi. A. (2008). Articulating a hermeneutic theory of writing assessment. Assessing Writing 13, 219–242

Polio, C., & Glewesl .M. (1996). Writing assessment prompts: how students choose. Journal of Second Language Writing,5(I),35-49.

Reinheimer. D. A. (2007). Validating placement: local means, multiple measures. Assessing Writing, 12 (2007) , 170–179

Saunders. P., & Scialfa. C.T. (2003). The effects of pre-exam instruction on students' performance on an effective writing exam. Written Communication, 20 (2), 195-212.

Schaefer, E. (2008). Rater bias patterns in an EFL writing assessment. Language Testing, 25(4): 465-493

Schils. E.D.J.,& van der Poel M.G.M., & Weltens. B. (1991). The reliability ritual. language testing. 8 (2), 125-138.

Schoonen,R. (2005).Generalizability of writing scores: an application of structural equation modeling. Language Testing, 22,1–30.

Shermis. M. D., & Long. S. K. (2009). Multitrait-multimethod analysis of fcat reading and writing: or is it writing and reading? Journal of Psychoeducational Assessment. 27 (4), 296-311.

Shi. L. (2001). Native-and nonnative-speaking EFL teachers'evaluation of Chinese students'English writing. Language Testing, 18(3):303-325

Shrum, J.L., Glisan, E.W. (2000). Teachers' Handbook : Contextualized Language Instruction. Foreign Language Teaching and Research Press : Beijing.

Singer, N.R., & LeMahieu, P. (2011). The effect of scoring order on the independence of holistic and analytic scores. The Journal of Writing Assassment. 4(1), December 2011.

Spandel, V. (2008). Creating Writers Through 6-Trait Writing Assessment and Instruction. 3rd ed. Pearson : *Allyn & Bacon*, Boston, United States.

Sudweeksa, R. R., & Reeveb. S., & Bradshawc, W. S. A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. Assessing Writing 9(2005): 239–261

Swain, M. (1993). Second language testing and second language acquisition:is there a conflict with traditional psychometrics? Language Testing, 10(2): 193-207

Tindal, G., & Parker, R. (1991). Identifying measures for evaluating written expression. Learning Disabilities Research and Practice, 6, 211-218.

Ur. P. (1996). A Course in Language Teaching: Practice and Theory. Cambridge University Press.

Weigle, S.C. (1999). Investigating rater/prompt interactions in writing assessment:quantitative and qualitative approaches. Assessing Writing 6(2), 145-1 78

Weigle, S.C. (2002). Assessing Writing. Cambridge University Press. Cambridge, UK.

Weigle, S.C., (2007). Teaching writing teachers about assessment. Journal of Second Language Writing 16, 194–209.

Weir, C. W. (2005). Language Testing and Validation: An Evidence-based Approach (Research and Practice in Applied Linguistics). Palgrave MacMillan, Basingstoke.

Williams.J.H.(1970). Testing the Writing Skills of Engineering and Science Students. Journal of Business Communication 8(1), 25-36

Williamson, M. M. (1993). Introduction to holistic scoring: The Social, Historical, and Theoretical Context for Writing Assessment. In M.M. Williamson and B. A. Huot (Eds.), Validating Holistic Scoring for Writing Assessment: Theoretical and Empirical Foundations (pp. 1-43). Cresskill, NJ: Hampton Press.

Williamson, M. M. (1994). Worship of efficiency: untangling theoretical and practical considerations in writing assessment. Assessing Writing, 1, 147-174.

Williamson, M., & Huot, B. (1992). Validating holistic scoring for writing assessment: theoretical and emperical foundations (written language). College Composition and Communication, 47, (4): 549-566.

Wilson, M. (2006). Rethinking Rubrics in Writing Assessment. Gloria Pipkin (Ed.). Heinemann Portsmouth,NH. USA.

Yancey,K.B.(1999).Looking back as We Look Forward:Historicizing Writing Assessment. College Composition and Communication,50 (1999): 483-503.

Yang, H.C. (2012). Modeling the relationships between test-taking strategies and test performance on a graph-writing task:Implications for EAP. English for Specific Purposes 31, 174–187.

Yu. G., Rea-Dickins. P., & Kiely. R. (2007). The cognitive processes of taking IELTS Academic Writing Task 1. IELTS Research Reports Volume 11. Retreived on Dec, 1, 2012 from www.ielts.org.