Porta Linguarum 27, enero 2017 93-106

Academic Writing Assessment: A Generic Encounter

Mohammad Aghajanzadeh Kiasi
Payame Noor University, Iran

Received: 1 November 2015 / Accepted: 16 May 2016
ISSN: 1697-7467

ABSTRACT: EFL/ESL researchers and teachers usually find the assessment of writing as a serious challenge, especially when several writing tasks are demanded in a curriculum. Having this in mind, this study aims to see through assessment priorities given to undergraduates' writing performances. For this aim, seventy Iranian university instructors with different teaching backgrounds were requested to assess two separate writing tasks entailing different rhetorical demands. Research findings disclosed important details of undifferentiated assessment approaches adopted by instructors. Despite a quantitatively significant difference in comments on *content richness* and leniency toward *discourse*, several writing constructs were accorded the same level of importance whatever the writing task type. This study casts fresh light on inadequacies in academic writing assessment and offers fruitful information to curriculum developers to consider a serious rethinking of language teachers education to sharpen teachers' assessment skills.
Keywords: content richness, discourse, generic assessment, task type, writing program.

Evaluación de la escritura académica: una experiencia genérica

RESUMEN: Los investigadores y profesores EFL/ESL ven la evaluación de la escritura como un reto, en particular cuando se exigen varias tareas en un currículum. Teniéndolo en cuenta, ese estudio apunta cómo establecer las prioridades de evaluación de las destrezas de escritura universitarias. Para alcanzar ese objetivo, se pidió a setenta profesores universitarios iraníes, con diferentes experiencias de enseñanza, evaluar dos tareas de escritura diferentes implicando varias preguntas retóricas. El resultado de la investigación reveló detalles importantes por lo que concierne a la aproximación indiferenciada de la evaluación adoptada por los profesores. No obstante se observó una diferencia cualitativa significativa en los comentarios sobre la riqueza del contenido y la indulgencia hacia el *discurso*, ya que se dio el mismo nivel de importancia a cualquier tipo de tarea escrita. Ese estudio arroja luz sobre la inadecuación de la evaluación de los escritos académicos y da información fructífera a los que redactan un currículum proponiendo una reflexión sobre la educación de los profesores de idiomas extranjeros, al fin de aclarar las habilidades de evaluación de los profesores.
Palabras claves: riqueza del contenido, discurso, evaluación genérica, tipo de tarea, programa de escritura.

1. Introduction

Written language is not restricted to the communication of information. It not only performs cognitive functions in explaining and supporting thought but can also be an essentially private medium to serve the functions of record keeping and storing both information and

literary works, hence transmitting culture. Inasmuch as these points are to be considered for pedagogy and assessment, students' compositions and written performance(s) cannot be assessed regardless of their purposes. In this sense, Hood (2010) claims that different genres are to be assessed differently and notes, since expositions are written, "to align readers to a point of view by arguments surrounding a thesis, they are intrinsically evaluative in their social purpose" (p. 7).

As a formative and inherently pedagogical endeavor, the assessment of students' writing processes and products is a key responsibility shouldered by mindful instructors considering the syllabus designed, the curriculum developed, and feedback provided for students' writings. Consequently, when instructors responsibly perform assessment according to the inherent demands of test items, they can be informed of their instruction effectiveness as it is reflected in their students' increasing proficiency and writing goals achievement. However, as Ferris and Hedgcock (2005) note, there are some contradictions which confront teachers while assessing students' writing performances. They believe the assessment of writing is frequently "framed in terms of institutionalized writing tests such as the essay component of the Test of English as a Foreign Language" which are product-oriented and do not take into account processes (pp. 300-301). Added to this problem is provision of formative feedback while assigning summative scores or grades, and these two objectives "may operate at cross-purposes" (Hedgcock & Lefkowitz, 1996, p. 288). Finally, audience under-representation is what has depicted many novice writers' performances as imperfect writings whose "audience is usually limited to the person (the teacher) who also designs, assigns, and assesses that writing" (Reid & Kroll, 1995, p. 18). All things considered, an investigation into assessment priorities given to writing variables or subscales while scoring can be of great importance. Thus, the present study explored how much university instructors' assessment priorities vary according to the writing type considered.

2. Literature review

2.1. EFL academic writing

A longitudinal critical evaluation of writing teaching programs in six EFL contexts, Japan, Poland, China, Germany, the USA, and Spain performed by Reichelt (2009) revealed most EFL contexts suffer from traditional approaches to teaching writing. The research findings pointed that language students at university level are being introduced into a narrow writing curriculum whose pedagogical features are limited as new concerns of writing such as discourse, voice, readership, and genres are unheeded and no assessment details were provided. The quality of language students' writing performances can also be linked to isolated grammar courses syllabuses designed for the first year of their university education. Schoonen, Snellings, Stevenson, and Van Gelderen (2009) propose the Inhibition Hypothesis, which predicts that the high demands of linguistic dimensions of FL composition will draw upon resources and inhibit attention to conceptual or ideological perspectives of FL writing, such as content elaboration, monitoring and higher-order revisions. Non-integrative grammar and product-oriented writing courses wherein the bare minimum of reading is deployed

aggravate this asymmetric attention paid to writing essentials (Hinkel, 2002). Given this critical investigation into several EFL contexts, each writing program inevitably incorporates a specific assessment approach to students' writings for accommodating its pre-determined demands. A traditional orientation toward teaching writing, alas, cannot be expected to be informed by modern assessment praxis.

2.2. Writing assessment challenges

Rezaei and Lovorn (2010) have shown that different interpretations of the scoring rubrics employed by different teachers give rise to an unreliable assessment which is to be tackled by developing locally appropriate rubrics mounted for specific groups of language learners. What should be added here is that no clear writing assessment theory has been advocated universally and as Cumming (2001, p. 214) asserts "there is no agreed upon definition on writing assessment" in L2 contexts. Hamp-Lyons (2003) finds this lack of consensus largely conditioned by the detachment of a writing assessment researcher from writers and class processes. Assessment initiatives are very occasionally negotiated or locally introduced. Instead, they are usually imposed from the top-down or determined by program outsiders who can ill afford to comply with local practitioners' expectations. O'Neill, Moore, and Huot (2009) believe that negative feelings about assessment can be intensified when program administrators are not familiar with possibilities for approaching large-scale assessment, as well as the key concepts, documented history, and recorded beliefs associated with various approaches. It is likely that many English teaching professionals do not have a clear understanding of the key concepts in educational measurement, such as validity and reliability, nor do they understand the statistical formulas associated with psychometrics. This unawareness very likely originates from "lack of communication between the composition community and the measurement communication" (Deborah, 2010, p. 257), which makes language teachers not foray into assessment details. What is more, many university English and composition instructors may shy away from finding validity theory and educational measurement as areas of interest or scholarship, desiring to focus instead on their own specific, immediate assessment needs or on what Fullan (1993) calls subjective reality according to which teachers work on their own perceived local concerns by making on-the-spot decisions, with little reference to assessment experts' propositions.

Moss (1994) and Huot (2002) have mainly reported writing assessment as a generic measurement activity which is rarely conditioned by context or task specificity by which the quality of a written task is to be assessed according to its inherent discoursal and functional specifications. They see academic writing assessment as an isolated activity to which a different agenda is attached by different teachers and experts and is largely detached from educational measurement. Hedgcock (2005) has come to the conclusion that "commonsense insights and criticisms, coupled with scrupulous empirical studies of numerous assessment variables, have led practitioners and researchers to raise serious concerns about both reliability and validity —particularly construct validity— in measuring L2 writing performance"(p. 607).

2.3. Scoring perceptions of task types

The rubric and prompt are two distinct features of each writing task that should be taken into account for assessment purposes. Douglas (2000) notes that the former illuminates

the specification of the objective, procedures for responding, the task format, and finally the evaluation criteria as they clarify the relative weighting attached to several features of writing. The prompt, however, acts as guiding language giving a vivid account of context. Therefore, the tasks or topics on which students are to write can exert considerable impacts on the content and its presentation in a written performance. He and Shi (2012) asserts that the quality and quantity of contents can vary inasmuch as topic familiarity can spawn richer arguments and general topics can bring about better organization and fewer linguistic errors. This issue may necessitate the development of appropriate prompts for class practices (He & Shi, 2012) or may be downplayed on the ground that prompt and task difficulty cannot significantly change scores and just socially demanding topics can pose a threat to scoring validity (Lim, 2010).

Cumming (2001) argues that instructors generally conceptualize ESL/EFL writing instruction in common ways but surprisingly their conceptualizations of student assessment vary by course type i.e. whether the courses are taught in reference to general or specific purposes for learning English. Unlike English for general purposes, the ESP courses writings are to be assessed specifically as to a clear expected function. Whatever the writing course, there is a natural tendency towards generic assessment of college writing to set common goals. Biggs (2003) is of the opinion that this assessment appeals to many instructors owing to its efficiency and simplicity that unite faculty around shared rhetorical orientations and writing purposes. In regard to the ease of application, there is very high likelihood that generic assessment takes priority over a localized or contextualized collegiate assessment framework which deals with idiosyncrasy of practices, purposes, majors, and courses. However, Anson, Dannels, Flash, and Housley Gaffney (2012) subscribe to the view that generic rubrics blur the contextual expectations by ignoring the discourse specificity of writing genres. They result in the creation of stereotypical classroom practices which best fit the single general framework or writing assessment.

A mixed-method research conducted by Moore (2015) has revealed that grammar and vocabulary are the most important writing constructs attended by raters for university entrance examinations in Japan but organization and content are the most serious assessment concerns for university writing classes, especially English for Academic Purposes. The study also lead to the conclusion that, in addition to these constructs, audience awareness and mechanics have to be considered for academic writing assessment. Investigating the qualities of EFL writing to which instructors specifically pay attention, de Haan and van Esch (2008) have found varying degrees of tendency to consider several writing constructs in students' performances. In particular, inconsistency in commenting on discourse competence and a particular focus on linguistic competence by raters participating in the study points to an imbalanced representation of writing features and lack of task-specific orientations in writing assessment.

Wolfe, Kao, and Ranney (1998) proposed the concept of *scoring focus* to suggest different scoring priorities that raters establish regarding their scoring competence. Similarly, Schaefer (2008) found out that the perceived importance of assessment criteria bears close relation to the severity and leniency raters show toward writing criteria. In a recent study undertaken by Eckes (2012) the raters' cognition and perceptions of writing scoring criteria of a large scale test (Test of German as a Foreign Language), was operationally investigated to suggest the likely tension between thinking and practice of writing assessors. He notes that more importantly perceived criteria are rated severely, and less valued features of writing

assessment are not seriously taken into account. In another survey through a questionnaire and follow-up interviews conducted by Ghanbari, Barati, and Moinzadeh (2012), intuition was believed to best suffice to assess Iranian students' writings inasmuch as "unmediated application of native rating scales would surface a hidden conflict between the assumptions behind these scales on the one hand and the realities of the local context on the other hand" (p. 85). Likewise, this current study was undertaken on academic writing with a focus on actual assessment priorities assigned while scoring.

3. Study

As claimed by Lumley (2002), although raters try to remain close to the rubric, they are heavily conditioned by the complex intuitive impression of the text gained when they first read it. Schoonen, Vergeer, and Eiting, (1997) believe that lay raters might be expected to be well capable of assessing certain qualities of a text (e.g., Content or Mechanics) reliably, because they exploit their strong intuitions or adequate knowledge about the aspect (subscale) in question. On the other hand, they may cling to their poorly formed intuitions and be incapable of doing so with other qualities (e.g., Usage or Style). Sifting through writing courses objectives and rubrics occasionally revisited by national curriculum developers, we can notice a dearth of formal, informed requests granted by faculties for having a major rethink of writing assessment performed in English writing programs at university level in Iran. On another reading, academic writing assessment in Iran has not received registered collective attention. Instead, feedback and responses provided by instructors during writing courses have shaped their own final assessment framework. Added to this issue is the paucity of information on academic writing scoring methods applied to several term-specific writing assignments. This caveat calls for an investigation into how Iranian instructors assess two commonly requested writing tasks at university level to which neither widely accepted assessment framework nor obligatory rubrics have been allotted.

3.1. Research question

This study was set out to investigate the current Iranian college instructors' realization of assessment of different writing tasks. Thus, the following research question was posed: Do college instructors apply task-specific criteria in the assessment of writing?

4. Method

4.1. Participants

Seventy university instructors from thirteen universities in Iran were asked, either by mail or in person, to assess two rhetorically different writings at a two-week interval. Two writings had been written by two female senior students of English, randomly selected from 19 willing volunteers. A lot of effort went into satisfying the requirement of subject representativeness through stratified sampling to proportionately include teaching contexts

(type of university wherein instructors were teaching such as state university, Islamic Azad university, and Payame Noor University) , ensuring the presence of key features of target population within the sample, especially instructors' educational background. In particular, Teaching English as a Foreign Language, English Literature, and Linguistics were the main academic fields of instructors whose teaching experience ranged between 1 and 18 years. Twenty of them had an M.A. degree varying in their fields (10 TEFL, 7 Linguistics, 1 English Literature, and 2 Translation Studies graduates), 28 instructors were PhD students (15, 7, and 6 students were doing their PhD, respectively in TEFL, Linguistics, and English Literature), and 22 participants had PhD varying in degree, namely TEFL (17) Linguistics (3), and English Literature (2).

4.2. Materials

Two writing tasks demanding two different types of written discourse were assigned to university instructors to be assessed. The first one was a letter, an expressive academic composition in which generally localized and situational conventions are demanded. The writing prompt was: *Write a motivation letter to highlight your qualifications to gain admission to university for a master degree program*. The second paper was an argumentative essay composed on this topic: *Substance use and oppositional Culture are the most serious triggering points for a rise in inner city crimes. Do you agree with this statement? Discuss far-reaching solutions to this dire predicament.* Raters followed a typical scoring scheme in Iranian universities by giving a score between zero and twenty accompanied by citing reasons for given scores or providing comments on the quality of the writings. Inter-coding reliability through Cohen's Kappa was applied to find out the degree of agreement between two coders. Unlike many studies (Brown, Glasswell & Harland, 2004; East, 2009; Wind & Engelhard, 2013) undertaken through predetermined marking schemes or questionnaires and interviews merely probing the assessment thinking, this study encompassed natural and intuitive assessment of the teachers to detect their genuine priorities given to two writing tasks in practice.

4.3. Procedure

The whole corpus of the current study was derived from comments on two writing performances, totaling 895 words phrased by seventy writing instructors. The comments were read and then coded by two expert coders whose decisions and labels attached to detected variables came under a close examination through Cohen's Kappa to find the degree of consistency. An interrater analysis $\kappa = 0.928$ deriving from .93 total agreement and .24 chance agreement suggested a high degree of agreement between two coders, ensuring the consistency of decisions. Overall, 620 writing subscales were identified and then classified into seven criteria: Grammar, Lexical knowledge, Coherence and Cohesion, Mechanics (spelling and punctuations), Content Richness, Discourse, and finally Format and Manner. As the last three mentioned constructs were expressed differently by different statements or phrases by instructors in their comments, some of them were asked to provide further elaboration to reach a consistent definition of reported constructs. To illustrate, the criterion *discourse* was assigned to comments such as specific language for a specific academic community or genre,

voice, and readership and *content richness* was commonly interpreted as a full coverage of the topic in terms of supporting ideas, statements, and evidence.

4.4. Data Analysis

To find out whether university writing instructors see two writing tasks differently from an assessment viewpoint a multivariate extension of McNemar's test, developed by Agresti, and Klingenberg (2006), was performed through Statistical Analysis Software (SAS). McNemar's test is a non-parametric method used on categorical data to figure out if the row and column marginal frequencies are equal. However, to square tables larger than 2x2, McNemar's generalization was employed. Statistically speaking, the bottom line of the study was to investigate the differences between paired vectors of binomial probabilities based on data from two dependent multivariate binary samples. Therefore, to find the degree of homogeneity in the marginal distributions a multivariate extension of McNemar's test was run.

5. Results

All seven variables coded and detected in two writings by inter-coders were firstly calculated by percentage, shown in Figure 1.

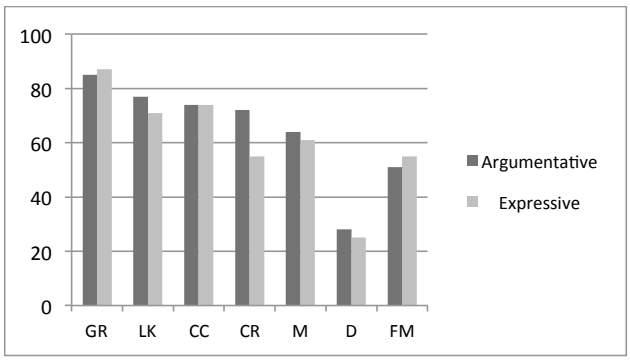

Fig. 1. A comparative presentation of commented variables

Note: Gr stands for *grammar*, LK for *lexical knowledge*, CC for *cohesion and coherence*, CR for *content richness*, M for *mechanics*, D for *discourse*, and MF for *format and manner*.

It can be clearly seen that *grammar* was the most frequently reported subscale unlike *discourse* which made up 25% and 28% of all raters' comments in the argumentative essay and the letter, respectively. Table 1 shows a breakdown on a 7x7 table for the frequency of reported writing constructs in two papers by seventy raters.

Table 1. *Computed frequency for constructs in two writings*

	Expressive							
Argumentative	GR	LK	CC	CR	M	D	MF	TOTAL
Grammar	121	115	113	112	106	81	97	745
Lexical Knowledge	110	104	102	101	95	70	86	668
Cohesion/Coherence	112	106	104	103	97	72	88	682
Content Richness	99	93	91	90	84	59	75	591
Mechanics	103	97	95	94	88	63	79	619
Discourse	78	72	70	69	63	38	54	444
Manner and Format	99	93	91	90	84	59	75	591
TOTAL	722	680	666	659	617	442	554	4340

The distribution of seven variables in two groups was investigated through Statistical Analysis Software (SAS) using Generalized Estimating Equation (GEE) (See Table2).

Table 2. Contrast results for GEE analysis

Contrast	DF	Chi-Square	p>ChiSq	Type
SMH	7	14.24	0.061	Score

All the test statistics and their respective p-value showed that there was a marginal homogeneity between reported writing constructs of an argumentative essay and those of a letter with $p< 0.05$ (See Table 3), i.e. the distribution of writing features considered for two different types of writing by 70 raters did not significantly differ. On another reading, instructors did not apply different assessment framework with different degrees of importance for different writing tasks. Inasmuch as *discourse* was the least reported criterion in writing assessment, and the widest gap was associated with *content richness* reported by raters in two papers several McNemar tests were applied to analyze the statistical association of each variable in two paired samples (See Table 3).

Table 3. *Crosstabulation and McNemar's Test results for Content Richness and Discourse*

	Essay	Letter		TOTAL	McNemar Test	
		YES	NO		Number of Valid Cases	Exact Sig.(2-sided)
CONTENT RICHNESS	YES	16	3	19	70	.008
	NO	15	36	51		
	TOTAL	31	39	70		
DISCOURSE	YES	48	0	48	70	.063
	NO	5	17	22		
	TOTAL	53	17	70		

The test yielded a *p*-value of .008<.05 which showed unequal probabilities of *content richness* in two types of writings. Simply put, *content richness* in two writings was differently perceived by academic writing raters. However, despite being the least reported assessment feature, *discourse* was equally discerned in two types of writings as the resulted p-value (.063) exceeded. It is worthy of note that except for *content richness* all assessment variables did not receive significantly different attention as to different rhetorical demands of two different writing tasks. Other five focused constructs, *grammar, lexical knowledge, mechanics, cohesion and coherence, and format and manner,* were also equally weighed in two writings as the result of their p-values exceeded .05.

6. DISCUSSION

Homogeneous comments provided for two different writing tasks show that Iranian academic writing instructors apply a single framework to two inherently different writing tasks. There is a mainstream assessment framework which may result from shared experiences of teaching similar syllabuses or fulfilling similar curricular expectations. An attempt to apply more generic assessment criteria is likely to violate the principle of *constructive alignment* (Biggs & Tang, 2007), which refers to the coherent relationship between specific learning goals, the methods of achieving those goals, and the assessment criteria used to judge the success rate. How much testing or assessing students' performances can affect subsequent

teaching and learning either positively or negatively is discussed through the issue of washback in applied linguistics. The positive effects are achieved when there is no mismatch between the content and format of the test and the content and format of the curriculum (Cheng and Curtis, 2008). Brown (2004) believes that a test having content validity shows its adherence to the curriculum and thereby sets the stage for the washback because careful preparation before the test and a thorough review of contents after the test can bring about positive outcomes for both students and teachers. However, what should be noted is that there can be false positive washback in a narrow curriculum in which students' achievements are assessed based on under-represented objectives or when the status quo is consciously or unconsciously being maintained. Given this situation, scant attention is likely to be paid to major writing concerns such as discourse, as revealed in this study. This collective leniency may also be an indirect result of the resource-demanding nature of linguistic processing in EFL compositions which hinders the opportunities for taking heed of higher-level concerns such as discourse, audience, arguments (Schoonen et. al, 2009). What is more, grammar and lexical knowledge were remarkably reported while assessing two writings in this study which shows a classical orientation toward these two important features. However, it will be more informative providing that the appropriacy of grammar and richness of vocabulary are taken into account. Comments on contextual specificity of these common features can raise students' consciousness and broaden their linguistic knowledge. To illustrate, instructors should be heedful of common initiation and termination of English letters with a present continuous structure, adjective or adverb embedded paragraphs for a descriptive performance, and lexical items such as *misdemeanor, sentence to, crack down, death penalty,* and *community service* for a composition on crime control. Finally, the only quantitatively significantly different reported assessment criterion in two types of writing was *content richness* in that its frequency while assessing the argumentative essay significantly outweighed that of the letter. Several probable reasons having potential for further investigations can be stated. Ideas, arguments, and concepts in an essay are better embraced as content, compared with fictionalized situations created by student writers in letter writing which may be handled as *format and manner*. Also, longer performances in essays (220-350 words) may simply establish a high expectation about topic coverage compared with laconic presentations of purposes in writing letters with a wording ranging from roughly 70 to 170 words.

Writing topics and tasks embody an inherent linguistic and discoursal degree of complexities. To illustrate, an argumentative task of a specific discourse community imposes more task difficulty and complexity than a descriptive text type on student writers and can also confront writing instructors both in instruction and testing (Crasnich & Lumbelli, 2004; Garate & Melero, 2004). Normally applied in EFL university writing sessions, general topics can ease the lexical and grammatical items selection as opposed to more specific topics demanding specific discourse, vocabularies, and community-based ideas and arguments that consequently lead to lower scores of university students (He & Shi, 2012). Thus, the prediction of mediocre performance of students in discourse –based writing which drives specific knowledge in terms of ideas and language in the absence of a broad writing curriculum can push teachers to more often apply general issues in writing assignments (Lim, 2010).

Bouwer, Béguin, Sanders, and van den Bergh (2015) emphasize the inclusion of multiple tasks in multiple genres of writing evaluated by multiple raters. Provided this is not feasible and economical, decisions should be restricted to genre-specific since the generalizability

of writing scores differs from genre to genre. Applying the same assessment framework to several writing assignments can give rise to validity concerns which are to be obviated through expert curricular decisions. O'Neill, Schendel, and Huot (2002) suggest that on this occasion instead of positioning a demand for assessment as an administrative task or an obligatory service assignment, the writing faculty and administration had better take part in occasionally communal action research, drawing on their expertise and education within a composition community to localize their curricular problems. In-house experientially designed rating scales make experienced and competent language teachers stand to benefit from intuition-based assessment as they are evolved and ameliorated over a period of time (Fulcher, 2003). Knoch (2007) finds the empirically developed version of a writing rating scale more promising as it results in higher reliability and validity. This thinking can be pursued through the socio-cognitive framework proposed by Shaw and Weir (2007) which explores validity in two phases of before (context validity and cognitive validity) and after the test event (scoring validity, consequential validity and criterion related validity). They find this framework a unified approach to the validity of assessment and argue that the validity of a writing test can be established when candidates are engaged in all the processing components described "as appropriate to the level of proficiency being assessed" (pp. 42-43). This framework has come highly recommended as Zainal (2012) suggests it for validation of classroom assessment in which controlled writing tasks and an analytic scale are applied to reach a unified ground in assessment and avoid different interpretations as are said to frequently occur in holistic scoring. Huot (1996) argues that in the writing assessment the validity should include a cogent theoretical foundation as well as empirical information regarding student writers' performances. He also adds that the purpose and use of writing should be clearly defined because if they are out of theoretical foundation of any writing program, they will cause an invalid measure of writing assessment.

7. Conclusion

The main concern of the study was to reflect Iranian academic writing instructors' actual assessment considerations investigated while assessing two rhetorically different writing tasks. A multivariate extension of MacNemar's test applied to seven criteria detected in two different written performances rated by seventy university lecturers and professors unveiled a generalized praxis; the more so because distinct writing tasks demanded within a classical university writing program were not allotted differentiated evaluative rubrics by university faculty. Also, *discourse* was poorly perceived or commented in writing assessment. In other words, voice, audience or readership, and language specificity for specific discourse communities were to a large extent ignored by college writing instructors. However, topic coverage or *content richness* in assessment of two investigated text types of writing in this study found significantly different degrees of favor with university lecturers and professors. This assessment feature received more significant attention in persuasive writings.

The insufficient attention directed to discourse in scoring academic writing and allocation of a single framework to assessment of two rhetorically different writing performances which can be the result of a narrow writing program in a university context can breed serious consequences. A task-specific scoring or commenting method, as a result, should be

assigned to separate university writing courses as they differ in demands and expectations. It means that task-specific explicit statements on assessment rubrics should be offered to link evidence to claims for each writing task. This assessment reform entails two key decisions taken by faculty: First, overall objectives of writing courses should be transparently established by program developers. Second, highly explicit witting goals should be set for each single genre of writing. Another considerable step which has to be made to initiate task-specific assessment of L2 writing is implementation of assessment training program. We should bear in mind that passing no comments on some criteria does not necessarily mean unawareness but maybe some discontinuity between instructors' response principles and their actual practices (Ferris, 2014). There is little doubt in the role raters training can serve in raising the scoring validity and reliability of writing assessment inasmuch as long teaching experience cannot necessarily bring about enhanced assessment skill and expertise and may appear as an obstacle when it may wrong obviate the need for being trained.

8. References

Agresti, A., & Klingenberg, B. (2006). "Multivariate extensions of McNemar's Test", in *Biometrics, 62,* 3: 921-928.

Anson, C. M., Dannels, D. P., Flash, P., & Housley Gaffney, A. L. (2012). "Big rubrics and weird genres: The futility of using generic assessment tools across diverse instructional contexts", in *Journal of Writing Assessment, 5,* 1, available from: http://www.journalofwritingassessment.org/article.php?article=57, accessed 15 February, 2014.

Biggs, J. B. (2003). *Teaching for Quality Learning at University*. Buckingham: Open University Press.

Biggs, J., & Tang, C. (2007). *Teaching for quality learning at university*. Maidenhead, UK: SRHE & Open University.

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

Brown, G., Glasswell, K., & Harland, D. (2004). "Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system", in *Assessing Writing, 9,* 2: 105–121.

Brown, H. D. (2004). *Language Assessment: Principles and Classroom Practices.* New York: Pearson/ Longman.

Cheng, L, & Curtis, A. (2008). "Washback or backwash: A review of the impact of testing on teaching and learning", in L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in Language Testing: Research Contexts and Methods* (pp. 3-17). New Jersey: Lawrence Erlbaum Associates.

Crasnich, S., & Lumbelli, L. (2004). "Improving argumentative writing by fostering argumentative speech", in G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), *Effective Learning and teaching of writing: A handbook of writing in education* (pp. 181-196). New York: Kluwer Academic Publishers.

Cumming, A. (2001). "ESL/EFL instructors' practices for writing assessment: Specific or general purposes?" In *Language Testing*, 182: 207-224.

Cumming, A. (2002). "Assessing L2 writing: Alternative constructs and ethical dilemmas". In *Assessing Writing, 8*: 73-83.

Deborah, C. (2010). "Assess thyself lest others assess thee", in T. Silva, & P. K. Matsuda (Eds.), *Practicing Theory in Second Language Writing* (pp. 245-262). West Lafayette, Parlor Press.

deHaan, P., & van Esch, K. (2008). "Measuring and assessing the development of foreign language writing competence." in *Porta Linguarum, 9*: 7-21.

Douglas, D. (2000). *Assessing Languages for Specific Purposes*. Cambridge: Cambridge University Press.

East, M. (2009). "Reliability and validity of rubrics for assessment through writing", in *Assessing Writing, 14*, 2: 88-115.

Eckes, T. (2012). "Operational rater types in writing assessment: Linking rater cognition to rater behavior", in *Language Assessment Quarterly*, *9*: 270–292.

Ferris, D. R. (2014). "Responding to student writing: Teachers' philosophies and practices", in *Assessing Writing, 19*: 6–23.

Ferris, D., & Hedgcock, J. (2005).*Teaching ESL Composition: Purpose, Process, and Practice*. London: Routledge.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson Longman.

Fullan, M. G. (2005). *The New Meaning of Educational Change.* London: Routledge Falmer.

Garate, M., & Melero, A. (2004). "Teaching how to write argumentative texts at primary schools", in G. Rijlaarsdam, H. Van den Bergh, & M. Couzijn (Eds.), Effective *Learning and Teaching of Writing: A Handbook of Writing in Education* (pp. 323-337). New York: Kluwer Academic Publishers.

Ghanbari, B., Barati, H., & Moinzadeh, M. (2012) "Problematizing rating scales in EFL academic writing assessment: Voices from Iranian context", in *English Language Teaching, 5*, 8: 76-90.

Hamp-Lyons, L. (2003). "Writing teachers as assessors of writing", in B. Kroll (Ed.), *Exploring the Dynamics of Second Language Writing* (pp. 162-189). New York: Cambridge University Press.

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

Hedgcock, J., &Lefkowitz, N. (1996). "Some input on input: Two analyses of student response to expert feedback on L2 writing", in *Modern Language Journal, 80:* 287-308.

Hedgcock, J. (2005). "Taking Stock of Research and Pedagogy in L2 Writing", in E. Hinkel (Ed.), *Handbook of Researching Second Language Teaching and Learning* (pp. 597-613). Mahwah, NJ: Lawrence Erlbaum Associates.

Hood, S. (2010). *Appraising Research: Evaluation in Academic Writing*. London: Palgrave.

Huot, B. (1996). "Towards a new theory of writing assessment", in *College Composition and Communication*, *47*: 549-566.

Huot, B. (2002). *(Re)Articulating Writing Assessment for Teaching and Learning.* Logan, Utah: Utah State UP.

Knoch, U. (2007). "Do empirically developed rating scales function differently to conventional scales for academic writing?", in *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*, 1: 1-36.

Lim, G. S. 2010. "Investigating prompt effects in writing performance assessment", in *Spaan Fellow Working Papers in Second or Foreign language Assessment, 8:* 95 - 116.

Lim, G. S. (2012). "The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters", in *Language Testing, 28,* 4: 543 –560.

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

Moore, Y. (2015). "Investigating valid constructs for writing tasks in EAP tests for use in Japanese university entrance examinations". British Council Report: Final report submitted

to the British Council, available from: https://www.britishcouncil.org/exam/aptis/research/publications/investigating-valid-constructs-writing-eap-tests.pdf, accessed 12 June, 2016.

Moss, P. (1994). "Can there be validity without reliability?" in *Educational Researcher, 23*, 2: 5–12.

Nelson, J. (1995). "Reading classrooms as texts: Exploring student writers' interpretive practices", in *College Composition and Communication, 46:* 411–429.

O'Neill, P., Schendel, E., & B. Huot. (2002). "Defining assessment as research: Moving from obligations to opportunities", in *Writing Program Administration, 26:* 10-26.

O'Neill, P., Moore, C., & Huot, B. (2009). *A guide to College Writing Assessment*. Logan, UT: Utah State University Press.

Reichelt, M. (2009). "A critical evaluation of writing teaching programmes in different foreign language settings", in R. M. Manchón (Ed.), *Writing in Foreign Language Contexts: Learning, Teaching, and Research,* (pp. 183-206). Clevedon, UK: Multilingual Matters.

Reid, J. M., & Kroll, B. (1995). "Designing and assessing effective classroom writing assignments for NES and ESL students", in *Journal of Second Language Writing, 4*: 17-41.

Rezaei, A. R., & Lovorn, M. (2010), "Reliability and validity of rubrics for assessment through writing", in *Assessing Writing* 15, 1: 18--39.

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

Shaw, S., & Weir, C. J. (2007). *Examining Writing in a Second Language, Studies in Language Testing 26*. Cambridge: Cambridge University Press/Cambridge ESOL.

Schoonen, P., Vergeer, M., & Eiting, M. (1997). "The assessment of writing ability: Expert readers versus lay readers", in *Language Testing, 14,* 2: 157-184.

Schoonen, R., Snellings, P., Stevenson, M., & van Gelderen. A. (2009). "Towards a blueprint of the foreign language writer: The linguistic and cognitive demands of foreign language *writing", in R. M. Manchón (Ed.), Writing in Foreign Language Contexts: Learning, Teaching, and Research* (pp. 77-101). Clevedon, UK: Multilingual Matters.

Wind, S., & Engelhard, G. (2013). "How invariant and accurate are domain ratings in writing assessment?" in *Assessing Writing, 18*, 4: 278–299.

Wolfe, E. W., Kao, C. W., & Ranney, M. (1998). "Cognitive differences in proficient and nonproficient essay scorers", in *Written Communication*, *15*: 465–492.

Zainal, A. (2012). "Validation of an ESL writing test in a Malaysian secondary school context", in *Assessing Writing, 17,* 1: 1-17.