TELL, Vol. 8, No. 1, Spring & Summer 2014, pp.1- 29

L2 pragmatics underrepresentation as a washback mechanism in EFL contexts: Linking test-design features to high-stakes test use

Azizullah Mirzaei [1]
Assistant Professor, Shahrekord University, Iran
Masoud Rahimi Domakani
Assistant Professor, Shahrekord University, Iran
Masoumeh Seyyed Rezaei
M.A. of TEFL, Shahrekord University, Iran

Received on May 1, 2014
Accepted on June 25, 2014

Abstract

Considering the status of 'washback' in validity theory and research, Messick warns that, to establish test validity, one should not rely merely on washback, with all its complexity, but should instead probe test design properties (i.e., authenticity and directness) and test-use characteristics likely to produce or intensify washback. Authenticity ensures that nothing essential is missing in the assessment of the focal construct, or minimal construct-underrepresentation. Similarly, directness deals with minimal construct-irrelevant variance. This study aimed to explore inadequate representation of L2 pragmatics in EFL curriculum and instruction of Iranian high schools as negative washback of construct-

[1] Corresponding author: Shahrekord University
Email Address: mirzaei-a@lit.sku.ac.ir

underrepresentation of this important dimension of communicative competence in the high-stakes National University Entrance Test (NUET). The participants were 100 EFL teachers, a sample of 220 pre-NUET students, and 50 post-NUET students in northwest, center, and southwest of Iran. A method-triangulation procedure was employed using questionnaires, observations, and semi-structured interviews. The results showed that the test's construct validity is compromised, meaning that NUET is too narrow and deficient, fails to sufficiently sample pragmatic competence as an important facet of 2discrete-point lexico-grammatical knowledge. Further, the findings demonstrated intense negative washback on L2 teaching-learning processes resulting from using NUET scores for drawing high-stakes action inferences in Iran. The findings suggest that, to facilitate optimal positive washback and adequate sampling of the criterion domain in L2 education, NUET must strive to minimize construct-underrepresentation and construct-irrelevant variance in design.

Keywords: Construct validity, washback, construct-underrepresentation, L2 pragmatics

1. Introduction

Second or foreign language (L2) learners' achievement of communicative intent in spontaneous interactions calls forth a repertoire of pragmatic knowledge defined as a set of internalized rules of how to use language in socioculturally appropriate ways in different contexts (Celce-Murcia & Olshtain, 2000). The role of pragmatic competence is crucial in the era of globalization where communication takes place across cultural boundaries as an everyday phenomenon (Taguchi, 2012) and its absence may have the speakers run the risk of appearing uncooperative or even rude and insulting. The importance of pragmatic knowledge derives from the fact that L2 learners in similar situations may resort to variable speech act patterns, forms, and semantic formulae to achieve the same communicative intents. Although some universal pragmatic knowledge may be shared across languages, there are noticeable differences in terms of appropriacy of the communicative acts produced by L2 learners and native speakers, which can result in

communication breakdowns. Several second language acquisition (SLA) researchers have recently incorporated pragmatic competence as the cornerstone of their proposed models of communicative competence which have greatly influenced the fields of language testing and pedagogy (e.g., Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980). Of particular importance to language testing has been the model of communicative language ability (CLA) proposed by Bachman (1990) and Bachman and Palmer (1996), which gives special momentum to pragmatic competence. Therefore, it is incumbent upon L2 learners to develop pragmatic knowledge along with other competences.

Research on L2 learners' performance in various target language contexts demonstrates that pragmatic competence does not develop automatically nor it accrues from learners' development of lexico-grammatical knowledge (Bardovi-Harlig, 1996; Eslami-Rasekh, 2005). Furthermore, a foreign language context provides little access to appropriate pragmatic input (Rose & Kasper, 2001), which implies that L2 pragmatics should be adequately represented within the instructional materials and activities of English-as-a-foreign-language (EFL) classrooms and also be sufficiently sampled and operationalized in current tests (Kasper & Rose, 2002; Roever, 2011). Although L2 practitioners have recently shown growing interest in teaching L2 pragmatics (Eslami-Rasekh, 2005), there has yet been no systematic planning to include pragmatics into mainstream classroom curriculum and instruction or, on a larger scale, into teacher education programs (Ishihara & Cohen, 2010). This problem arises largely from the fact that the construct of pragmatic competence is gravely underrepresented in both the high-stakes (proficiency and gate-keeping) tests (Grabowsky, 2008; Roever, 2011) and the classroom assessment (Cohen, 2010), perhaps, due to the variable nature of pragmatic behavior and the inherent complexities involved in its assessment (Eslami & Mirzaei, 2012). Washback research has shown that testing, especially if it creates winner-loser, success-failure, and rejection-acceptance, comes before the teaching and learning process (Cheng, 1997; Shohamy, 2001) and test content directly affects the quality and

type of L2 teaching-learning practices in language classrooms at schools (Chapman & Snyder, 2000; Wall, 2005). Therefore, it can be hypothesized (as this study does) that roots of minimal attention to L2 pragmatics in textbooks, classrooms, and instructions (e.g., in Iran) may be traced to a corresponding construct-underrepresentation in the relevant testing policies or systems.

On the other hand, if any high-stakes test is found to induce in the education system curricular and instructional changes that impede the development of the cognitive or communicative skills that the test was originally designed to measure, it lacks 'systemic validity' (i.e., having optimal positive effects on L2 teaching and learning) which is an important facet of construct validity (Messick, 1996). In the case of NUET in Iran, the test should include authentic and direct samples of the pragmatic behaviors, besides other communicative behaviors, of the language being learnt, and there should be little difference between activities involved in learning the language (e.g., social interaction) and activities involved in preparing for the test (e.g., test wiseness strategies). This study focused on the systemic validity of NUET and attempted to probe whether the test suffers from construct-underrepresentation of pragmatic ability and construct-irrelevance variance. Further, the study aimed to unearth if lack of authenticity or directness in NUET engenders negative washback in EFL curriculum and instruction across the country.

2. Literature Review

The reciprocity between teaching and testing, or 'curriculum alignment' (Cheng, 1997), has recently been reflected in terms of 'washback hypothesis' (Alderson & Wall, 1993) and 'systemic' or 'consequential' validity (Messick, 1996) in language teaching and testing literature. In essence, washback hypothesis assumes that teachers and learners do things they would not necessarily otherwise do because of the test, and that tests are powerful determiners of what happens in classrooms. Tests influence what, how, rate, and depth of in-class teaching and learning (Alderson &

Wall, 1993). In simple terms, not only do 'good' tests encourage the use of good teaching-learning processes, but also they are more or less directly useable as teaching-learning activities (Pearson, 1988). Therefore, tests offer effective mechanisms for 'bias for best' or 'work for washback' (Swain, 1985, pp. 42-44) and levers for promoting curricular innovations (Stecher, Chun, & Barron, 2004). In practice, however, most tests have over time become associated with high stakes, in terms of rationing future admission or employment opportunities (Chapman & Snyder, 2000), and, in turn, led to unintended negative consequences for different stakeholders and to unbeneficial washback for teaching-learning processes (Hawkey, 2006). Several researchers in the field of educational measurement have related washback to test validity through using notions such as 'washback validity' (Morrow, 1986) or 'systemic validity' (Frederiksen & Collins, 1989) and have argued that, to ensure validity, a test should demonstrate beneficial washback. In other words, the test should promote curricular and instructional changes that foster the development of the cognitive skills that the test is designed to measure.

As noted, Messick (1996) examines waskback as an instance of the consequential aspect of construct validity—treated as an integrative, unified concept—and attempts to link positive washback to authenticity and directness properties of valid tests (or assessments). Messick makes a distinction between test washback *per se* (i.e., practically accrued from test interpretation and use) and the (positive or negative) effects of good or bad educational practices regardless of the test quality. That is, good teaching or learning might exist in the case of a poor test, or vice versa, poor teaching and learning in the case of using a construct-validated test. He argues, however, that this is highly circumstantial and dependent upon other things done in the educational system (e.g., the good or poor educational practices) than the test use. Therefore, evidence of washback can be claimed only if a logical and evidential link can be forged between the teaching-learning processes (or outcomes) and the test properties (e.g., authenticity and directness) likely to have influenced them. In plain terms, negative washback *per se* should be associated (only) with the introduction and use of less valid tests

and positive washback with the introduction and use of more valid tests. In a similar vein, Alderson and Wall (1993) note that washback is a complex phenomenon and cannot be directly related to a test's validity. Rather, one should focus on test properties of authenticity and directness that are likely to generate washback and examine their implications for test validity (Messick, 1996).

'Authentic' and 'direct' assessment tasks, in the case of language testing, involve realistic simulations or criterion samples so that the available time, resources, and processes to perform them (approximately) resemble those in the real world. The major concern of authenticity (in language assessments) is to ensure that no important component of the focal construct i.e., communicative language ability) is left out, that is, minimal construct-underrepresentation. On the other hand, the directness issue seeks to minimize constraints on examinee behavior associated with construct-irrelevant method variance such as structured item forms, restrictive response formats, test wiseness, and differential guessing tendencies. Authenticity and directness can never be fulfilled in ideal forms in instructional and assessment settings. Nonetheless, these favorable properties of assessment foster positive washback, whereas construct-underrepresentation and construct-irrelevant variance generate negative washback and thus pose threats to validity (Messick, 1996).

The notions of construct-underrepresentation and construct-irrelevant variance have recently received increasing attention (e.g., Downing, 2002; Haladyna & Downing, 2004; McNamara, 2006). Haladyna and Downing (2004), for instance, relating construct-irrelevant variance to construct validity in high-stakes testing, argue that various psychological and situational sources of construct-irrelevant variance might contaminate the measurement by inducing systematic error variance in test scores. Downing (2002) probed construct-underrepresentation in achievement tests in medical education and found that most tests were too short and used inadequate items that tapped low levels of cognitive domain such as recall or recognition of facts and resulted in teaching-to-the-test behaviors. Research on the implications of the notions for test

validity in language testing is, however, scarce. Xi (2010) relates these notions to test fairness (or 'comparable validity') and holds that these test characteristics should have no systematic and appreciable effects on test score interpretation and use for relevant groups of examinees. Following Messick's conception of washback mechanisms, Green (2006) explicates a predicative model of test washback in which he makes a link between test design features, test use, and observed educational practices. He relates design features to contexts of test use and notes that washback will be most intense where the test results are seen as important and associated with high-stakes decisions, such as university entrance. Nonetheless, Wall (2005) and Green (2006) indicate that the washback dynamics in terms of the relationships between test design features, test use, and classroom behaviors are under-researched and most of this work just takes the form of recommendations to test developers.

As one of the integral components of recent communicative models of language ability, pragmatic competence still remains a neglected area in educational contexts in many respects in comparison to other aspects of learners' development of CLA, such as knowledge of grammar and vocabulary. Specifically, research on assessment of L2 learners' pragmatic has received too little attention and emphasis, even far less than research on interlanguage pragmatics (Ishihara & Cohen, 2010; Roever, 2011). Pragmatics assessment instruments still have not found their way to standard decision-making testing practices in (non-)academic contexts. In addition, even current tests of L2 pragmatics have been criticized for relying on outdated or inauthentic methods, under-representing the construct, sampling observations too narrowly, and lacking an explicit interpretive argument (Grabowsky, 2008). Eslami and Mirzaei (2012) argue that the reason for lack of attention to assessing L2 pragmatics may be related to the complexities involved in the task, for instance, tension between authenticity and practicality, indivisibility of sociopragmatic and pragmalinguistic facets of pragmatic competence, and considerable variability of pragmatic norms across situations or individuals. The lack of generally-accepted, systematic measures to assess learners'

pragmatic knowledge (Bachman, 1990) and the minor role that pragmatics assessment, despites its enormous importance, plays in high-stakes testing situations have in turn led to widespread negligence of interlanguage pragmatics in mainstream L2 teaching and learning.

Despite the inherent complexity and lack of systematicity in approaching the task, questions remain whether the enormity of the task should instill trepidation and deter (even seasoned) researchers and test developers from exploring its dimensions. Further, if current language testing systems go on under-representing pragmatic competence in nowadays' high-stakes tests, will they not compromise the construct validity of their tests? In simple terms, as far as validity theory requires test developers to adequately define the test construct in terms of which the test score is said to have meaning (McNamara, 2012), won't such test scores potentially be misleading indicators of their focal construct? How would the negative washback resulted from this high-stakes test interpretation and use be reflected in educational practices of teachers and learners? And finally, will they stay accountable for the undesirable consequences of their actions to teaching-learning processes in L2 classrooms? These and other concerns pertinent to construct-underrepresentation of L2 pragmatics in high-stakes testing practice in Iran constitute the theme of this study.

Secretariat of Higher Council of Education in Iran (2006) defines the main pedagogical goal of EFL instruction at high schools as to "enable students to use at least one (foreign) language to communicate with others" (p. 43), implying that students after graduation should be able to use English appropriately in social-academic contexts. However, experience has shown Iranian EFL learners (in high school and even at graduate levels) achieve less (especially pragmatic) success in L2 communication. One of the effective ways to minimize pragmatic failure is developing instructional materials and activities to engage students into meaningful communications and explicitly raise their consciousness of cross-cultural differences and gaps in their pragmatic knowledge (Eslami-Rasekh, 2005; Ishihara, 2010). The current study primarily

intends to examine if L2 classroom materials or instructions are planned in a way to make high-school students aware of pragmatic conventions essential to L2 communication success. Second, it explores whether pragmatic dimension is under-represented in the focal construct of the high-stakes NUET and, finally, whether (or how) this test-design feature is reflected as unbeneficial washback in EFL teachers' and learners' educational practices.

3. Method

3.1 Participants and Settings

The participants comprised a total sample of 395 EFL teachers, high school students, and undergraduate students in northwest, center, southwest, and south of Iran. They were aware that they were being observed or surveyed to gather data for a study endorsed by university and Education Ministry, but had no knowledge of the real research goals involved. From them, 25 students and teachers participated at the pilot-testing of the instruments, and the remaining 370 others (i.e., 100 EFL teachers, 220 high school pre-NUETs, and 50 post-NUETs) answered the developed questionnaires. The EFL-teacher sample comprised 73 female and 27 male pre-university and third-grade high-school teachers in Tehran, Tabriz, Urmia, Shahrekord, Isfahan, and Shiraz. As to their academic degrees, 84 held B.A. in English language or literature and 16 had M.A. in TEFL (Teaching English as a Foreign Language). Their teaching experience ranged from 1 to 40 years (M = 16) and their age-range was 24-68 (M = 40). The high-school students comprised 142 pre-university and 78 third-grade students with the age-range of 16-19 studying Mathematics, Natural Sciences, Humanities, and Art. The post-NUET students were 50 freshmen (aged 16 to 19) who had just passed NUET and entered university. They were selected from different major universities in the aforementioned parts of the country. Finally, 25 EFL teachers and 38 high-school students were randomly selected to participate in the semi-structured interviews. As noted above, all the

participants gave their consent for the current study, and their anonymity was ensured.

Despite the communicative goals claimed for L2 education in Iran, EFL learning is generally carried out for formal, academic purposes giving considerable momentum to lexico-grammatical knowledge mainly gained through reading or translation skills. Iranian students study EFL for three years at junior high school, three years at senior high school, and one year at pre-university level before sitting for NUET and entering university. NUET is a high-stakes norm-referenced test which is developed and administered at the end of each academic year to pre-university students of different majors. NUET for each 'group' (or major) of candidates is modular encompassing main 'general' and 'technical' subject matters.

3.2 Instruments

Two parallel questionnaires—EFL teachers' questionnaire (TQ) and pre-NUET students' questionnaire (SQ)—were developed to assess, first, the participants' perceptions of representation of L2 pragmatics in classroom teaching-learning practices or instructional materials. Second, if it is underrepresented, the questionnaires also examined the extent to which this phenomenon is the undesirable washback of construct-underrepresentation in the high-stakes test of NUET in Iran. Each questionnaire comprised 20 items in the form of two subscales, namely, pragmatic-construct underrepresentation (PCU) and washback (WS) scales. A four-point Likert-scale ranging from never (1) to always (4) was used to assess the participants' responses. In addition, an independent pragmatics underrepresentation questionnaire (UQ) was developed for assessing post-NUET undergraduate students' perceptions of construct-representation (of L2 pragmatic competence) in NUET. The UQ comprised of 10 Liker-type items with four anchors ranging from 'no item' (1) to 'a lot of items' (4).

Specifically, the questionnaires were constructed based on the related literature of the CLA model (Bachman, 1990; Bachman &

Palmer, 1996) and test washback (e.g., Alderson & Wall, 1993; Cheng, 2005; Green, 2006; Wall, 2005). Experts' judgments also ensured that the content and method of the instruments are sufficiently representative and practical. Pragmatic competence was operationalized in terms of knowledge of speech acts or functions; sensitivity to language varieties, register, or naturalness; and knowledge of figurative or cultural concepts. The instruments were translated into Persian to account for the confounding variable of participants' differential L2 knowledge and increase the validity and reliability of the responses. Moreover, back-translation and pilot-testing were used to ensure precision, ease and time of administration, and face validity. An adjunct part eliciting participants' demographic information was also included. In brief, the questionnaires were then subjected to Principal Component Analysis for examining the underlying factor pattern(s), factor-related loadings of items, and assuring construct validity. Prior to performing PCA, factorability of the data was ensured. Cronbach's alpha coefficient was used to estimate the internal consistency of the questionnaires which resulted in acceptable values: TQ = 0.79, SQ = 0.84, and UQ = 0.81.

Furthermore, an observation checklist was developed to estimate the representation of pragmatic ability or skills in L2 classrooms. The checklist included and drew upon Bachman's CLA model to tap the core features of pragmatic competence using a four-point Likert-scale. Finally, three open-ended questions were formulated as the interview script focusing on (i) the types of L2 materials and instructional practices implemented and the goals stressed in L2 classrooms, (ii) whether they served L2 learners' pragmatic needs, and (iii) whether construct under-representation in NUET had any negative washback on this.

4. Results

A quantitative-qualitative analysis method was employed arguing that reducing test design features, teaching-learning processes, and washback mechanisms merely to (quantified) tables or figures would sacrifice the whole for only a partial picture. Descriptive

statistics and chi-square tests were calculated for the questionnaire data to see how the participants perceived the extent of construct-underrepresentation in NUET and its negative washback on their educational practices. Following this, the findings of L2-classroom observations, the NUET-layout examination, and the semi-structured interviews are presented.

The descriptive statistics (in Table 1) suggest that both EFL teachers and high-school students believed that the current L2 teaching-learning processes do not adequately attend to L2 pragmatic competence in high-school EFL classrooms. Similarly, post-NUET students believed that NUET was not designed in a way that could assess L2 learners' pragmatic awareness and competence to use English appropriately in social situations. EFL teachers and pre-NUET students also believed that the lack of sufficient sampling of pragmatic domain in NUET's focal construct had considerable negative washback on the underrepresentation of L2 pragmatics in EFL classrooms in Iran.

Table 1: Descriptive statistics for TQ, SQ, and UQ

Instruments		Participants	N	Min	Max	Mean	SD	Skewness	Kurtosis
Pragmatics in L2 Classroom	PCU	EFL teachers	100	1	3.2	2.33	.75	.21	-.36
		Pre-NUETs	220	1	3.5	1.88	.96	.76	-.41
	WS	EFL teachers	100	1.9	4	2.74	.87	-1.01	-.62
		Pre-NUETs	220	1	3.9	2.56	1.09	-.81	-1.27
UQ		Post- NUETs	50	1	2.3	1.35	.53	1.5	2

4.1 L2 Pragmatics in EFL Curriculum and Instruction

A set of chi-square tests was run on the PCU data of both TQ and SQ to examine the extent to which instructional materials and educational practices are conducive to interlanguage pragmatic development. Pragmatic facets operationalized in the PCU scale were (direct and indirect) speech acts, language functions (ideational, manipulative, and heuristic functions), dialect and variety, register (subject matter of language use, field of discourse,

and style of discourse), and sensitivity to naturalness (idiomatic expressions and routines). Table 2 displays the chi-square results.

Table 2: Chi-Square results for pragmatic representation

Pragmatic competence	**P**	**N**	**S**	**O**	**A**	X^2	*p*
Speech acts	T	23.0%	39.0%	32.0%	6.0%	63.98	.00
	S	66.5%	14.8%	14.3%	4.3%	827.69	.00
Ideational functions	T	11.0%	63.0%	21.0%	5.0%	125.87	.00
	S	55.9%	26.4%	12.3%	5.5%	556.81	.00
Regulatory & Instrumental	T	10.0%	56.0%	30.0%	4.0%	97.2	.00
	S	57.0%	23.0%	14.8%	5.2%	579.85	.00
Interpersonal	T	14.0%	38.0%	36.0%	16.0%	38.55	.00
	S	54.3%	23.5%	15.7%	6.5%	549.18	.00
Heuristic & Ideational	T	7.0%	41.0%	36.0%	16.0%	38.55	.00
	S	33.5%	31.7%	23.5%	11.3%	207.33	.00
Dialect/Variety	T	15.0%	39.0%	33.0%	13.0%	39.07	.00
	S	31.3%	25.7%	27.8%	15.2%	142.75	.00
Register	T	18.0%	51.0%	21.0%	10.0%	79.65	.00
	S	37.0%	36.1%	19.6%	7.4%	261.2	.00
Style of Discourse	T	16.0%	51.0%	25.0%	8.0%	78.08	.00
	S	39.6%	29.6%	22.6%	8.3%	271.17	.00
Naturalness	T	17.0%	45.0%	29.0%	9.0%	60.2	.00
	S	33.0%	27.8%	27.8%	11.3%	172.94	.00
Cultural referents	T	18.0%	37.0%	36.0%	9.0%	46.07	.00
	S	46.1%	22.2%	18.7%	13.0%	335.39	.00

P (Participants) T (Teachers) S (Students) N (Never) S (Sometimes) O (Often) A (Always)

The chi-square results for EFL teachers and pre-NUET students were statistically significant indicating noticeable underrepresentation of almost all aspects of pragmatic competence in the teaching-learning processes. Simply put, EFL teachers in the current educational system are not much concerned with raising (or enhancing) high-school students' awareness of different facets of pragmatic competence (essential to L2 communication success) through planned instruction. High-school students also felt that the employed instructional materials and activities are not aimed to foster their interlanguage pragmatic ability to fulfill their

communicative intents in real-life situations in an accurate and appropriate manner. It can be concluded that L2 pragmatics is seriously underrepresented and, by way of illustration, the Cinderella of EFL classrooms in Iran.

4.2 Representation of L2 Pragmatics in NUET

Another set of chi-square tests was run on the post-NUET survey data to probe whether NUET sufficiently represented pragmatic competence based on post-NUET examinees' perceptions of the test content. The chi-square results (Table 3) were all statistically significant, meaning that NUET items, from a post-NUET perspective, underrepresented different facets of L2 pragmatics and over-emphasized other (lexico-grammatical) aspects of the CLA.

Table 3: Chi-Square results for pragmatics representation in NUET

Pragmatic competence	No items	A few items	considerable items	A lot of items	*X2*	*p*
Speech acts	60.0%	38.0%	2.0%	0.0%	79.70	.00
Ideational functions	47.0%	22.0%	4.0%	0.0%	121.70	.00
Heuristic & Ideational	40.0%	42.0%	16.0%	2.0%	78.41	.00
Dialect/Variety	62.0%	28.0%	4.0%	6.0%	162.51	.00
Naturalness	80.0%	18.0%	2.0%	0.0%	146.90	.00
Cultural referents	100%	0.0%	0.0%	0.0%	82.81	.00

4.3 NUET Analysis

Closer inspection of the NUET content and method indicated that all NUET versions for different 'groups' of candidates, namely, Natural Sciences, Mathematics, Social Sciences, and Art, strictly followed an underlying test-specification blueprint in terms of test sections, number of items for each section, and test method (i.e., item and response formats). All NUET versions comprised four

major sections (i.e., vocabulary, grammar, cloze passages, and reading comprehension) and employed structured item forms and fixed response formats, that is, multiple-choice items. In other words, the test conformed closely to what is generally referred to as system-referenced indirect test model in which pragmatic competence has no place. Table 4 compares the extent of representation for each language component and indicates that reading skill and lexical knowledge received the most prominence. Grammatical knowledge was the next important component on the testing (and teaching) agenda, and still most blanks in cloze passages tapped textual and lexico-grammatical knowledge at the expense of other essential CLA components, most importantly, pragmatic competence.

Table 4: Representation of CLA components in NUET

Test Items	Vocabulary	Grammar	Cloze	Reading Comprehension
Representation	27.2%	16.8%	20%	36%

The following recently-used NUET-items illustrate the focal construct of interest, which is (mainly) L2 lexico-grammatical knowledge, and the preferred test method, i.e., fixed response format. A likely interpretation is that the test mostly taps the candidate's mastery of language as a 'system' largely through de-contextualized test items and is thus running contrary to the principles and practices of communicative language teaching and testing.

Natural Sciences Group (NUET, 2010)

Part A: Grammar and Vocabulary

77. Most students are studying hard ------- prepare themselves for their exams.
1) so as 2) so that 3) in order to 4) in order that

80. “Scientists are trying to find out when an earthquake occurs.”
“Occur” means ------.
1) include 2) continue 3) produce 4) happen

4.4 Perceptions of NUET Washback

Additional chi-square tests were run on the WS section of TQ and SQ to explore participants' perceptions of negative washback resulting from construct-underrepresentation and high-stakes test use on the lack of emphasis on pragmatics in L2 classrooms. Different related facets of the WS scale were: (unbeneficial) washback on negligence of L2 pragmatic features in teaching materials, classroom testing practices, and instructional processes; under-development of pragmatic and communication skills; lack of sensitivity to L2 sociopragmatic conventions in spoken and written language use; and inadequate coverage of L2 cultural references and figurative speech. The chi-square results for all the levels of washback on teaching-learning processes (Table 5) were statistically significant for both EFL teachers and pre-NUET students, meaning that the current negligence of instructional pragmatics stems from a parallel underrepresentation of pragmatic competence in NUET since its birth decades ago. Further, the fact that NUET use and score interpretations were associated with high stakes for candidates has in turn intensified the unfavorable washback. Teachers see no reason why they should invest in (pre-NUET) students' interlanguage pragmatic development as the ultimate L2 educational goals seem to be developing learners' knowledge of (morphosyntactic) L2 system, working with the L2 as a way to improve their cognitive, academic skills, and increasing their testwiseness for better future test performance. Another reason may simply be that L2 pragmatics is highly underrepresented in currently-used coursebooks.

Table 5: Chi-Square results for washback analysis

Washback	P	N	S	O	A	*X2*	*p*
Materials	T	2.0%	19.0%	45.0%	34.0%	14.85	.00
	S	7.7%	25.0%	38.0%	29.3%	35.00	.00
Educational practices	T	2.0%	38.0%	36.0%	24.0%	30.20	.00
	S	14.1%	18.6%	34.1%	33.2%	70.00	.00
Testing practices	T	0.0%	16.0%	45.0%	39.0%	6.49	.03
	S	10.9%	15.5%	47.7%	25.9%	65.75	.00
Grammar-Instruction	T	9.0%	28.0%	41.0%	22.0%	20.60	.00
	S	8.6%	26.8%	30.5%	34.1%	30.12	.00
Vocabulary-Instruction	T	4.0%	15.0%	42.0%	39.0%	49.83	.00
	S	5.9%	18.2%	36.8%	39.1%	63.33	.00
Testing pragmatics	T	3.0%	24.0%	40.0%	33.0%	68.55	.00
	S	12.3%	14.5%	37.3%	35.9%	18.20	.00
Analyzing Discourse	T	14.0%	7.0%	31.0%	47.0%	28.18	.00
	S	14.5%	10.0%	44.1%	31.4%	79.20	.00
Appropriacy	T	12.0%	14.0%	31.0%	43.0%	12.65	.00
	S	8.0%	28.2%	27.3%	36.5%	43.04	.00
Figurative language	T	11.0%	16.0%	31.0%	41.0%	32.68	.00
	S	13.2%	29.5%	26.4%	30.9%	66.43	.00
Cultural referents	T	3.0%	10.0%	42.0%	35.0%	15.90	.00
	S	13.2%	17.3%	25.5%	34.1%	136.48	.00

4.5 Classroom Observations

Percentages were obtained for the amount of time EFL teachers and students were observed to be devoting to different dimensions of pragmatic competence (Table 6).

Table 6: Underrepresentation of pragmatics in L2 classroom

Component	Checklist Items	N	S	O	A
Speech acts	1	100%	0.0%	0.0%	0.0%
Language functions	2-10	90.8%	6.8%	2.4%	0.0%
Dialect/Variety	11	60.0%	32.0%	8.0%	0.0%
Register	12-16	77.0%	18.0%	5.0%	0.0%
Naturalness	17	68.0%	32.0%	0.0%	0.0%
Cultural/Figurative reference	18	72.0%	28.0%	0.0%	0.0%

The observation results further indicated that pragmatics received almost the least attention in the L2 classrooms. No explicit efforts were made to address how to recognize or produce contextually-appropriate speech acts and language functions in various situations. EFL teachers conducted few metapragmatic discussions or awareness-raising activities (e.g., watching a video clip of a complaint-apology situation and doing appropriacy-oriented reflections) to enhance target-pragmatic-features' saliency. Nor did they engage students in typical pragmatic-oriented tasks such as role-playing or simulations to raise their capability of functioning in a competent manner in real-life situations of L2 use. No attention was made to sociopragmatic constraints and cultural references of L2 speech community to empower students to interpret interlocutors' intentions in L2 communication. Also, there was little exposure to authentic audio-video input so as to increase students' sensitivity to natural speech routines in daily communication. On the other hand, analyzing field-notes revealed that both teachers and students were hard pressed to do reading, vocabulary, and grammar exercises of the coursebooks. Occasionally, NUET-like test items were introduced and examined

in terms of the lexico-grammatical and test-strategy issues involved. In most cases, L1 was the medium of instruction, and (silent) reading comprehension was sometimes emphasized.

4.6 Semi-Structured Interview Results

Scrutiny of the audio-recorded interviews yielded emic-oriented insights into the washback mechanisms at work in L2 classrooms and how minimal representation of essential CLA components in high-stakes testing system can lead to imbalanced emphasis on L2 lexico-grammatical knowledge and reading (through translation). In the following quote from a female EFL teacher in Tabriz (northwest of Iran), this issue is clearly pointed out:

> *... Students read and translate reading passages into L1. We teach L2 aspects tested in NUET because passing NUET is the major educational goal of our students…They are not interested to be engaged in L2 pragmatic-oriented practices. They even protest when I speak English in class and prefer to understand everything through translation. They only aim is to pass the examinations and NUET. So, I prefer to reinforce the vocabulary and grammar knowledge.*

She then explains how the test-polluted context pushes practitioners to keep up with the market (or test industry) that has grown around this high-stakes testing situation. The fundamental goal of 'commodities' at this market is, obviously, not to foster students' CLA as this is not a concern in NUET, but rather to increase (L2 system-oriented) testwiseness:

> *... I choose a series of multiple-choice tests for each lesson from Gaj, Ghalamchi, etc. publishers and share them with the students as assignments…. We try to use these market provided 'test books' as supplementary to ensure better achievement in NUET.*

Two other comments below illustrate 'excessive coaching for exams' and negligence of L2 pragmatics as negative washback reflected in classroom practices. EFL teacher from Shahrekord (southwest):

> *Teaching materials are nothing special except for books, chalk, and board. Most of the time we work on 'test books' like Khat-e-Sefid or*

> *Kanun (containing past NUET tests and similar mock questions and practices), and the major focus is on reading comprehension, vocabulary, and grammar. ... Last-year high-school students prefer to learn how to tackle the tests in NUET. Their goal is simply NUET; therefore, they are not so attentive to learn L2 pragmatics nor interested in learning to speak or negotiate meaning in L2.*

EFL teacher from Tehran (capital):

> *We are mostly pressed for time to finish the coursebook, and thus attending to other L2 aspects not covered in the books is a matter of time. ... Right now, our purpose is the book, final exams, and NUET. If textbooks and NUET actually assessed other L2 aspects such as pragmatic competence, teachers would absolutely allot sufficient time to engage in related instructional practices too.*

5. Discussion

The results of the current study indicated that L2 pragmatics was highly underrepresented in the high-stakes (university-entrance) testing system and L2 classrooms in Iranian educational context. Particularly, it was evidenced that L2 teaching-learning processes put great momentum on students' achievement of lexico-grammatical knowledge and their ability to read or comprehend L2 texts mainly through translation. In contrast, little attention was devoted to students' development of pragmatic competence as an essential component of CLA (Bachman, 1990). The results raised grave concerns as mainstream instructional materials and practices attached no considerable importance to students' sensitivity to appropriacy norms in communicative language use in real-life or simulated contexts and to cross-cultural differences. Furthermore, considering that tests have enormous power over what takes place in the classroom (Alderson & Wall, 1993), the NUET testing system seriously underrepresents different (illocutionary-sociolinguistic) dimensions of pragmatic competence, gives undue weight to L2 usage and specific reading sub-skills, and thus results in extreme construct-irrelevant variance. More importantly, this

evidenced minimal authenticity and directness in the design of NUET encouraged non-communicative teaching-learning processes in L2 classrooms (besides its other social, individual impacts that were of no concern in this study), which can compromise the test's 'systemic,' or on a larger scale, construct validity (Messick, 1996).

In practice, the existing design and method shortcomings of NUET, on one hand, and the high-stakes NUET-score interpretation and use, on the other, produce unfavorable, intense washback effects on EFL teaching and learning at large and, according to Messick (1996), pose threats to test validity. Apart from the obvious underrepresentation of other important dimensions of the criterion performance such as listening and speaking, NUET makes inadequate sampling of the criterion domain, defined as "the relevant domain of behavior, knowledge, or skills in relation to which we need to establish the candidate's standing" (McNamara, 2006, p. 33). Complementary NUET analysis in this study demonstrated that the test contains few items which are designed to tap examinees' knowledge of pragmalinguistic or sociopragmatic facets of communicative language use. Furthermore, as noted earlier, NUET adopts structured item forms or restrictive response formats which procedurally generate considerable construct-irrelevant method variance such as testwiseness (in coping with various item forms) and differential guessing tendencies (towards multiple-choice items). Therefore, the test is too narrow as it fails to include important dimensions of the focal construct and, at the same time, is too broad since it contains excess reliable variance that is irrelevant to the interpreted construct (Messick, 1996).

In addition, NUET test use and score-based interpretations play a major role in Iranian educational system in making important decisions about candidates, for instance, admission to higher education and, in turn, attainment of differentially-lucrative careers in future. High-stakes test use and score-based interpretations constitute the next crucial linkage in the chain of washback mechanisms as it binds test design features to washback or other consequences of test use in Messickian conceptualization of validity (Bachman, 2005; Green, 2006). Bachman (2005) addresses this lacuna (test use) in validity research by proposing an 'assessment

use argument' framework to link assessment qualities such as validity, usefulness, and fairness to score-based inferences and consequences of test use for stakeholders. Similarly, Green (2006) draws test design and test use together in his 'predictive model of test washback' and notes that while test design issues (e.g., construct-representation) are associated with the direction of washback (positive or negative), test use features are closely related to washback intensity. NUET washback is thus most intense since stakeholders see the test results as important and associated with high-stakes decisions, such as university entrance and job opportunities. It was evidenced that EFL teachers tailored their teaching content to the test content (mainly NUET and final exams), overemphasized system-referenced lexico-grammatical dimensions of L2 knowledge and reading skills that are well-represented in NUET, and overlooked L2 pragmatic facets that are not adequately covered in NUET no matter how much essential to CLA they are. Meanwhile, students seemed to have been treated as passive recipients of the content knowledge (represented in textbooks and summative tests) which was supposed to be extensively regurgitated later for subsequent test delivery. Most interviewees believed that, in the test-driven context of NUET, students' real language needs, learning interests, and communicative intentions are generally neglected in the instruction.

In broad terms, similar results have been obtained by other NUET-related washback studies that point to the negative washback of the test on curriculum, language teachers, and learners. Ghorbani (2008), for instance, found that NUET results in excessive coaching (in terms of class time) for the test. Further, Mirzaei and Roshani (2011) found that NUET use and score-based interpretations have created a rift between EFL teachers' selected-teaching-styles in classrooms (to teach the textbook and test content) and students' preferred-learning-styles (to learn EFL for communicative purposes). Based on their results, whereas students preferred to learn English through collaborative, hands-on activities (e.g., role-playing, going to language labs, or using chat-rooms), their EFL teachers rarely did so and mainly employed individual-teaching-

activities (e.g., silent and then loud reading of passages, translating the texts, and teacher-initiated question-answer) that could serve their test-preparation purposes. Most of the participants agreed that they selected such instructional processes since they felt what is important in the current educational milieu is 'to teach the book to the test.' Finally, Riazi and Razavipour (2011) found evidence of unbeneficial washback of (centralized summative) testing practice in Iran (e.g., NUET) on EFL teachers' 'agency' in the sense that they are pushed by the current testing system and the stakes involved to reproduce teaching-to-test behaviors.

To sum up, the evidenced construct-underrepresentation in NUET, its overemphasis on lexico-grammatical knowledge and reading skills using restrictive response formats, and high-stakes test use have collectively resulted in intense negative washback on L2 teaching-learning processes. Attempts were made to establish washback by linking test-use consequences to test design features through the linkage of high-stakes decisions made based on test-score interpretations as conceptualized by Messick (1996), Bachman (2005), and Green (2006). This approach was taken since "it is problematic to claim evidence of washback if a logical and evidential link cannot be forged between the teaching and learning outcomes and the test properties thought to influence them" (Messick, 1996, p. 247). It can be argued that the negative-washback evidence documented here and in other NUET-related studies in the literature does not contribute satisfactorily to the test's 'systemic validity' or the consequential aspect of construct validity. In terms of the theoretical rationale underlying the test design, this ongoing underrepresentation of L2 pragmatics hardly renders NUET a CLA-criterion sample and thus precipitates the bad, inefficient educational practices currently employed in L2 classrooms.

The test also uses structured item forms and restrictive response formats and seldom involves realistic simulations or samples of criterion domain. Therefore, the tasks, processes, and available time or resources do not parallel those in the real world. Further, the test stresses knowledge of grammar, vocabulary, and reading skills to the detriment of pragmalinguistic and sociopragmatic facets of

communicative competence. In practice, these sources of invalidity generate considerable construct-irrelevant variance in test scores. For instance, candidates' test scores may differ remarkably, not due to their differential attainment of CLA, but rather owing to their varying level of test wiseness and test-preparation (e.g., Xie, 2013), or how prepared they are to tackle the specific test method at hand. Moreover, imbalanced emphasis on construct-irrelevant lexico-grammatical facets of L2 knowledge can have teachers or learners pay undue attention to overcoming the construct-irrelevant difficulty as opposed to fostering communicative proficiency (Messick, 1996). In sum, the test stops short of fulfilling optimal authenticity and directness as the touchstone criterion of its evidential basis, is not 'fit for purpose', and keeps on producing invalid test scores and inferences.

6. Conclusion

The primary concern of the current study was to probe underrepresentation of L2 pragmatics in EFL curriculum and instruction in Iranian high schools as a negative washback accruing from a parallel construct-underrepresentation and construct-irrelevant variance (as test design properties) in the NUET testing system and as a consequence of high-stakes test-score interpretations. The study thus explored washback mechanisms by linking NUET test-design features and high-stakes test use to teaching-learning processes in L2 classrooms across the country, following Messickian conception of construct validity, Bachman's (2005) 'assessment use argument,' and Green's (2006) 'predictive model of test washback.' The findings revealed that L2 pragmatics is highly underrepresented in both the focal construct of NUET and L2 instruction. More importantly, it was found that high-stakes test use acted as a linkage in the washback chain to intensify unbeneficial effects of the test on the teaching and learning of L2 pragmatics as a crucial dimension of CLA. In other words, since the test use is associated with negative washback, it can 'distort the curriculum' and influence the attitudes, behavior, and motivation of

teachers, learners, and parents (Alderson & Wall, 1993). Furthermore, it was indicated that the NUET testing system, because of its reliance on structured item forms and restrictive response formats as well as its overemphasis on lexico-grammatical knowledge and reading sub-skills, causes significant construct-irrelevant variance (e.g., testwisenees) in the candidates' performance and test scores and, before everything, makes stakeholders pay undue attention to surmounting construct-irrelevant difficulty to the detriment of fostering interlanguage pragmatic competence. It was then argued that the testing system lacks even minimal levels of authenticity and directness in its design and the evidential and consequential bases do not support the test use.

This study can be of substantial contribution to the under-visited domain of exploring washback mechanisms in terms of test design properties (i.e., authenticity and directness) and test use and score-based interpretations from a Messickian perspective. The study underscores Messick's position that, to establish validity, related research should not merely rely on washback, with all its complexity and variability, but rather should turn to test in-built properties and test use characteristics likely to produce or intensify washback and consider what they might mean in validity terms (Messick, 1996). In other words, washback evidence should be sought in light of an argument of evidential and consequential bases for particular test use and score interpretations. Therefore, distinction should be made between behavioral and attitudinal changes in teachers and learners that are evidentially and consequentially linked to the introduction and use of important tests and good or bad instructional practices that are, in essence, effects of other forces operative on the educational scene. Furthermore, the scope adopted here can be of interest to those L2 researchers and practitioners who pursue instructional pragmatics and may wonder why teaching and assessing L2 pragmatics, despite being of prime importance, still have not found their way to L2 classrooms and (communicative) proficiency measures.

References

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics, 14*(2), 115-129.

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

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bardovi-Harlig, K. (1996). Pragmatics and language teaching: Bringing pragmatics and pedagogy together. In L. F. Boutan (Ed.), *Pragmatics and language learning* (pp. 21-39). Urbana: University of Illinois.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics 1*(1), 34-56.

Celce-Murcia, M., & Olshtain, E. (2000). *Discourse and context in language teaching*. Cambridge: Cambridge University Press.

Chapman, D. W., & Snyder, C. W. J. (2000). Can high stakes national testing improve instruction: Reexamining conventional wisdom. *International Journal of Educational Development, 20*(6), 457-474.

Cheng, L. (1997). How does washback influence teaching? Implications for Hong Kong. *Language and Education, 11*(1), 38-54.

Cheng, L. (2005). Changing language teaching through language testing: A washback study. *Studies in Language Testing*, *21*. Cambridge: Cambridge University Press.

Cohen, A. (2010). Approaches to assessing pragmatic ability. In N. Ishihara & A. Cohen (Eds.), *Teaching and learning pragmatics: Where language and culture meet* (pp. 265-285). London: Longman.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant

variance and construct underrepresentation. *Advances in Health Sciences Education, 7*(3), 235-241.

Eslami-Rasekh, Z. (2005). Raising the pragmatic awareness of language learners. *ELT Journal, 59*(3), 199-208.

Eslami, Z. R., & Mirzaei, A. (2012). Assessment of second language pragmatics. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoynoff (Eds.), *The Cambridge guide to second language assessment* (pp. 198-208). Cambridge: Cambridge University Press.

Frederiksen, J.R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18*(9), 27-32.

Ghorbani, M. R. (2008). The washback effect of the university entrance examination on Iranian English teachers' curricular planning and instruction. *The Iranian EFL Journal, 2*, 60-87.

Grabowsky, K. (2008). Measuring pragmatic knowledge: Issues of construct underrepresentation or labeling? *Language Assessment Quarterly, 5*(2), 154-159.

Green, A. (2006). Watching for Washback: Observing the Influence of the International English Language Testing System Academic Writing Test in the Classroom. *Language Assessment Quarterly, 3*(4), 333-368.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant varriance in high-stakes testing. *Educational Measurement: Issues and Practice*, *23*(1), 17-26.

Hawkey, R. (2006). Impact theory and practics. *Studies in Language Testing*, 24. Cambridge: Cambridge University Press.

Ishihara, N. (2010). *Where does instructional pragmatics fit and to what extent?: Teacher development and L2 pragmatics.* Paper presented at the Pragmatics and Language Learning Conference, Kobe, Japan.

Ishihara, N., & Cohen, A. (2010). *Teaching and learning pragmatics: Where language and culture meet*. London: Longman.

Kasper, G., & Rose K. R. (2002). *Pragmatic Development in a Second Language*. Oxford: Blackwell Publication.

McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy, *Language Assessment Quarterly, 3*(1), 31-51.

McNamara, T. (2012). Language assessments as shibboleths: A poststructuralist perspective. *Applied Linguistics, 33*(5), 564-581.

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

Mirzaei, A., & Roshani, N. (2011). A critical study of the impact of high-stakes testing practice on the teaching-learning process. *English Language Assessment, 6*(December),75-102.

Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing* (pp. 1-13). London: NFER/Nelson.

Pearson, I. (1988). Tests as levers for change. In D. Chamberlain & R. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation,* (pp. 98-107). ELT Document 128. London: Modern English Publications in association.

Riazi, M., & Razavipour, K. (2011). (In) agency of EFL teachers under the negative backwash effect of centralized tests. *International Journal of Language Studies, 5*(2), 123-142.

Roever, C. (2011). Testing of second language pragmatics: Past and future. *Language Testing, 28*(4), 463-481.

Rose, K., & Kasper, G. (2001). *Pragmatics in language teaching*. Cambridge: Cambridge University Press.

Secretariat of Higher Council of Education in Iran (2006). *Collection of regulations by the higher council of education*. Tehran: Madrese.

Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Longman.

Stecher, B., Chun, T., & Barron, S. (2004). The effects of assessment-driven reform on the teaching of writing in Washington State. In L. Cheng & Y. Watanabe (Eds.), *Washback in language testing: Research contexts and methods* (pp. 53-71). Mahwah, NJ: lawrence Erlbaum Associates.

Swain, M. (1985). Large-scale communicative testing. In Y. P. Lee, C. Y. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35-46). Hong Kong: Pergamon Press.

Taguchi, N. (2012). *Context, individual differences and pragmatic competence*. London: Multilingual Matters.

Wall, D. (2005). The impact of high-stakes examination on classroom teaching: A case study using insights from testing and innovation theory. *Studies in Language Testing*, *22*(3). Cambridge: Cambridge University Press.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing, 27*(2) 147-170.

Xie, Q. (2013). Does test preparation work? Implication for score validity. *Language Assessment Quarterly, 10*(2) 196-218.