Journal of Teaching Language Skills (JTLS)
36(4), Winter 2018, pp. 171-201- ISSN: 2008-8191
DOI: 10.22099/jtls.2018.28165.2440

The Structural Invariance of a Model of Washback to Test Takers' Perceptions and Preparation: The Moderating Role of Institutions

Kioumars Razavipour *
Assistant Professor
Shahid Chamran University of Ahvaz
razavipur57@gmail.com

Zohreh Gooniband Shooshtari
Assistant Professor
Shahid Chamran University of Ahvaz
zshooshtari@yahoo.com

Mahboobeh Mansoori
M.A.
Shahid Chamran University of Ahvaz
sweetiearad88@yahoo.com

Abstract

Test washback is held to be complicated and multifaceted in that a host of cultural, social, individual, test, and institutional factors are involved in shaping it. Thus far, the majority of washback studies have had as their focus the role of teachers in test washback or washback to teachers. How educational environments or institutions might function in isolation or in interaction with other factors in shaping washback to the learners and test takers has not received adequate research attention. The current study examined the mediatory role of academic institutions in washback to learners' perceptions of test content and test preparation. To this aim, 86 senior English students from two universities, one a top tier and the other a low tier one, completed two questionnaires: one on test takers' preparation practices including test analysis, test taking skills, drilling target skills, and socio-affective strategies; and the other on test takers' construal of test demands and uses as well as their expectation of success on the test. The data analyzed through partial least squares structural equation modeling revealed that a washback model based on expectancy-value theory explains a moderate amount of variance in test preparation. Further, for test takers from the low tier university, favorable perceptions of test content were associated with more value placed on test taking. However, Multi-group analysis pointed to group-invariance of the model across the two institutions,

Received: 21/01/2018 Accepted: 08/05/2018
* Corresponding author

indicating a lack of strong evidence for the mediatory role of educational environments in washback to test takers' perceptions and preparation.

Keywords: washback, structural invariance, institution, MALT, test perceptions, test content, test use

Test washback has garnered much prominence in language testing literature over the last few decades. Defined as the effect of testing on learning and teaching (Alderson & Wall, 1993), washback has come to the forefront, partly thanks to Messick's unified matrix of construct validity, which posited that consequences of tests constitute an aspect of construct validity (Messick, 1989). The existing literature has it that washback is not a linear process from the test to education; instead, it is a complex phenomenon mediated and moderated by numerous individual, political, cultural, educational, and institutional factors (Shohamy, Donitsa-Schmidt, & Ferman, 1996; Spratt, 2005; Wall & Alderson, 1993; Watanabe, 2004b). When it comes to factors mediating the test washback, the focus has thus far been mainly on either test factors ((Shohamy, Donitsa-Schmidt, & Ferman, 1996) or teacher factors (Alderson & Hamp-Lyons, 1996; Watanabe, 1996). How test washback might vary across institutions has not been the subject of empirical research to date. More specifically, how test takers' perceptions of test content and their preparation practices are affected by test washback differently across academic institutions remains unexplored.

There is substantial evidence suggesting that learners' perceptions of educational environments are more powerful determinants of their academic attainment than the actual instruction taking place (Entwistle, 1991). This is consistent with a phenomenological view of human perceptions (Mackey, 2010). As such, educational institutions are likely to foster different perceptions of the elements of education: teachers, textbooks, and examinations. Perceptions of assessments may be seen as an element of conceptions of the educational environment. Thus, it is justified to speculate that different educational environments likely induce different conceptions of and reactions to assessments. Cheng,

Sun, and Ma (2015) hypothesize that "students' perceptions of tests are likely to be shaped by the school context, for example, by their teachers and peers" (p.446). Similarly, Xie and Andrews (2013) asserted that test washback is more intense to learners in lower tier universities. Although the noted claims made by Cheng et al. (2015) and Xie and Andrews are plausible, there is little, if any, empirical evidence that this is, in fact, the case. Informed by this gap, the current study investigates possible differential washback from M.A Language Test (MALT) to test takers from two different tertiary education institutions: a top-tier university and a low tier one. To this end, we compared the two groups of test takers' perceptions of MALT's content and their test preparation practices. This is a worthwhile question that if interaction effects between test takers' characteristics and those of institutions do in fact exist, it would mean that uniform policies and procedures aimed at fostering beneficial washback (e.g., Bailey, 1996; Brown, 2005) in specific and educational reform at large should be reconsidered and tailored in accordance with the particulars of institutional contexts. In other words, in addition to learner and teacher factors, institutional characteristics should also enter the complex equation of test washback and hence reform and program evaluation.

Review of Literature

Before Wall and Alderson (1993), washback was perceived to be of a deterministic nature. That is, it was believed that tests bring about specific changes to learning and teaching and it was, and still is, this understanding of test washback that led many policymakers to use tests as levers of change in educational systems (Andrews, 2004). It was soon realized; however, that test washback is embedded in a complex network of social, cultural, and personal factors; hence, is a highly complex phenomenon. To entangle this complexity, Watanabe (2004) proposed a model of test washback comprising of *aspects*, *dimensions*, and *mediating factors*. Five sets of variables were subsumed under the mediating factors component: test factors, prestige factors, macro-context

factors, personal factors, and micro-context factors. As this study is about the latter two sets of variables, we limit this review to studies with a similar focus. Also excluded from this review are studies dealing with teachers' characteristics in mediating washback. Accordingly, in the remaining of this section, we review the existing research on washback to test takers and the role that micro-context factors play in shaping washback.

Compared to washback to teachers, research into washback to the test takers and learners is rather sparse (David, 2016; Qin, 2011; Xie & Andrews, 2013). Test takers' characteristics that have been studied concerning washback include motivation, perceptions, and conceptions of design and uses of assessment, as well as their assessment literacy and socio-economic background. The test taker characteristic that has been subject to considerable research is motivation. To investigate how test taker motivation plays out in washback, scholars have drawn on various motivational theories in psychology such as possible self-theories (Zhan & Andrews, 2014), expectancy-value theory (Xie & Andrews, 2013), and achievement goal theory. Xie and Andrews (2013) investigated how perceptions of test design and uses affected the preparation practices Chinese test takers employed in preparing for College English Test (CET). It was found that instrumental motivation and favorable attitudes toward the test content were associated with intense test preparation. Using structural equation modeling, they found that a washback model postulated based on expectancy-value theory was of adequate model fit to the data. Kaur, Noman, and Awang-Hashim (2017) found that students with mastery goals and those with performance goals differed in their reasons for test preparation, in their attitudes towards test taking, and in their preferred mode of test preparation. In an experimental study, Smith, Worsfold, Davies, Fisher, and McPhail (2013) found that having students apply testing criteria increases their assessment literacy and higher levels of assessment literacy were, in turn, appeared to be a good predictor of variation in test scores.

Though individual learner factors do matter in explaining washback, it must be remembered that learning often takes place, or at least triggered, within the context of an educational environment. In general education, there is an established strand of research on school climate (Freiberg, 1999), a review of which goes beyond the scope of this paper. The National School Climate Center (NSCC) gives the following definition for school climate

> School climate is based on patterns of students', parents', and school personnel's experience of school life and reflects **norms**, **goals**, **values**, interpersonal relationships, **teaching and learning practices**, and organizational structures. A sustainable, favorable school climate fosters youth development and **learning** necessary for a productive, contributing, and satisfying life in a democratic society (emphasis added, 2014, para.3).

The above definition contains several keywords that are of immediate relevance to test taking and test preparation. Test takers' goals, values, and their learning, as well as learning for test taking, are all embedded within the norms, goals, and values of the school climate. Both teaching and learning practices are also considered as dimensions of school climate. The implication is that to understand test preparation better, the psychological tradition of studying test preparation as an individual practice should be complemented with broader social, and institutional perspectives.

Nevertheless, whether and how educational environments moderate test washback has received little if any, research attention. In designing their study, Xie and Andrews (2013) alluded to the role of educational environment in test washback. Because low tier universities spend more resources on test preparation, Xie and Andrews (2013) speculated that washback from the CET exam is more intense in lower tier universities. On that speculation, they chose their study sample from a lower tier university. They did not deem it necessary to provide any evidence that this was, in fact, the case: that the same test produces differential

washback across tertiary institutions of different academic prestige. Likewise, Watanabe (2004) maintains that the school setting where test preparation takes place mediates washback. None of the noted sources provide any evidence regarding how institutional context bears on washback to test takers.

The current study constitutes an attempt to empirically put to the test the mediatory role of institutions on students' perceptions of tests and how those perceptions are in turn reflected in their test preparation practices. More specifically, we aim to examine how test takers from two higher education institutions hold various perceptions of test uses, test values, test design, and expectation of success on the test. Moreover, how such perceptions translate into test preparation strategies will also be investigated. To that aim, the following research questions guided the study.

1. To what extent does a washback model consisting of perceptions and values predict test preparation for MALT?
2. Do test takers from the top and low tier universities differ in their perceptions and preparation for MALT?
3. Does a model of washback show model-invariance between test takers from low and top-tier universities?

Given the complexity of washback and the many factors that are possibly in interaction (Watanabe, 2004), studies accommodating multiple factors hold the potential to predict and explain the washback mechanism. In this study, following Xie and Andrews, we postulated a theoretical model based on expectancy-value (EV) theory and tested against empirical data. However, unlike Xie and Andrews, we were primarily interested not in the model per se but model invariance across test takers from two higher education institutions. To explain how the conceptual model of washback makes sense in the context of MALT, a brief review of selected components of the EV theory is in order.

EV theory holds that faced with a task human beings ask themselves two fundamental questions: Do I want to do it? Moreover, Can I do it?

(Wigfield & Eccles, 2000). The first question corresponds to the value component and the second to the expectancy component of the theory. However, the answers they would give to these questions hinge on some precedent factors. In the first place, their short and long-term goals in life bear directly on whether they want to do a task. Besides, individuals' understanding of what it takes to do a task would affect the answer they give to the expectancy question; whether they believe they can do it or not. The logic of the current study is that the answer test takers give to each of the noted questions might not be solely determined by their characteristics but also by the climate of the academic institution in which they live their educational life. Theoretically speaking, the present study adds a social dimension to the somewhat psychological expectancy-value theory.

Methods

Setting and Participants

Admission to Iran's higher education at national universities is controlled by a standardized, nation-wide, multiple-choice exam, administered once a year during late April. Very high stakes are attached to the exam, as nearly a million candidates compete every year for admission into graduate programs at national universities. The fact that candidates far outnumber the available seats at universities has increased the stakes of the test. For entry into Masters' English language program, candidates should take a test that has locally come to be known as M.A Language Test (MALT, henceforth). MALT consists of a general English proficiency module and a specific module, with the latter varying across the three orientations of Translation Studies, English Literature, and English Language Teaching. The focus of this study is the former component, comprising of reading comprehension, grammar, vocabulary, and a set of cloze passages. The test is designed, administered, and scored by the national organization of educational testing (NOET).

The participants of this study were 86 primary English students, 50 females and 36 males, preparing for the MALT at the time of this study.

As most undergraduate students in English language departments across the country begin to prepare for MALT during their fourth year in college, we approached senior students from two different higher education institutions in Khuzestan, Iran.

Forty-nine participants were from Shahid Chamran University (SCU), and 37 from Jahad Daneshgahi University (JDU). The two universities, SCU and JDU, differ in their prestige as well as in their national and international rankings. SCU is a major national top-tier university while DJU is among the low tier universities (see www.isc.ir). The two also differ in their student population size and their history. Formerly known as Jondishapur, SCU's establishment dates back to the third century (Farhady & Hedayati, 2009). In contrast, JDU is a rather young tertiary education institution established less than three decades ago.

Regarding the adequacy of the two samples used, we should note that Partial Least Squares Structural Equation Modeling (PLS-SEM), the approach used in this study, is known for its resilience with small sample sizes and deviations from normal distribution; yet, there are minimum requirements that have to be met (J. Hair, Sarstedt, Hopkins, & Kuppelwieser, 2014). Hair et al. (2016) recommend that sample size should be equal to or larger than the larger of the following two indexes: "ten times the largest number of formative indicators used to measure one construct; or ten times the largest number of inner model paths directed at a particular construct in the inner model" (p. 109). Given that the two indexes were equal in the current study (see Figures 1), the minimum required sample size is forty participants. By this standard, the SCU sample exceeds the minimum requirement whereas the JDU sample is just slightly below the requirement. Given the size of the JDU population of senior English significant students, this was inevitable.

Instrumentation

To capture test-takers' perceptions and preparation behaviors, this study utilized two self-report, Likert scale questionnaires. We intended to use the very instruments Xie and Andrewes used but realized that College English Test in China differed in content, format, and function from MALT. For one thing, in contrast to MALT, which is an admission test, CET was an exit test for all undergraduate students from diverse fields of study. Moreover, the questionnaires Xie and Andrews used captured all the four language skills, whereas MALT does not have any oral or productive components. These differences led us to eliminate many items and add new ones to the two questionnaires.

The test perception questionnaire captured the following four subscales on a six-point Likert scale (see Table 1): *perceptions of test content* (six items), *perceptions of test use* (3 items), *the expectation of success* (four items), and *test value* (two items). *Perception of test content* captured participants' beliefs about knowledge and skills crucial for taking MALT (e.g., I must grasp the main idea instantly in the reading passages). *Perceptions of test use* elicited participants' goals for taking MALT (e.g., I take MALT to get an M.A degree for job seeking). *The expectation of success* asked participants about how confident they were about their success on MALT (e.g., how do you predict your overall performance on MALT). Finally, the *test value* subscale asked the participants how important they consider MALT (e.g., Performing well on MALT will be useful for my future).

Table 1.

Components and Number of Items in the Perceptions Questionnaire

Construct	*Number of items*	*Maximum possible score*
Perceptions of test use (PTU)	3	18
Perceptions of test content (PTC)	6	36
Expectation of success (ES)	4	24
Test value (TV)	2	12
Total	15	80

The second questionnaire (see Table 2) measured test takers' test preparation practices on a five-point Likert scale. This scale consisted of four subscales: *test analysis* with 13 items (e.g., I spend more time on my weak points, I analyze MALT past test papers to identify frequently assessed points), *rehearsing test-taking skills* with 13 items (e.g., I choose options through logic elimination), *drilling target skills* consisting of 12 items (e.g., I keep on reading English newspapers/websites; I listened to English audio files), and *socio-affective strategies* with 8 items (e.g., I tried to learn from others). We elaborate on the reliability and validity of the instruments in the results section where measurement models are evaluated.

Table 2.
Components and Number of Items in the Preparation Questionnaire

Construct	*Number of items*	*Maximum possible score*
Test analysis	13	65
Test-taking skills	13	65
Drilling target skills	12	60
Socio-affective strategies	8	40
Total	46	230

Data Analysis

As stated earlier, one of the dual aims of this study was to examine the adequacy of a theoretical model of washback across two groups of test takers set apart by their institutional affiliation. For a conceptual model to enjoy generalizability across contexts and populations, evidence regarding model invariance must be provided (Byrne, 2010; Shin, 2005). To examine the equivalence of the washback model across institutions, we used SPSS and Multi-Group Partial Least Squares structural equation modeling (PLS-SEM), using SmartPLS, version 3. Given our sample size, PLS-SEM was preferred to covariance-based SEM in that PLS-SEM is almost free of the daunting assumptions, such as the need for large samples of participants and strict data normality requirements,

which one has to meet in covariance-based SEM (Ravand & Baghaei, 2016). In the context of PLS-SEM, this is accomplished via the bootstrapping, which "is a type of robust statistic that simulates how resampling would replicate a study from a population" (LaFlair, Egbert, & Plonsky, 2016, p.46). Another reason behind using PLS-SEM was that PLS-SEM is more appropriate in domains where theory is not well developed (Garson, 2016). Except for the study by Xi and Andrews (2013), we were aware of no studies putting the model to empirical test.

One additional advantage of PLS-SEM is its capacity to handle both reflectively measured and formatively measured constructs in the model (J. F. Hair, Hult, Ringle, & Sarstedt, 2016). The difference between the two types of constructs is that in reflectively measured constructs, which is the standard and default understanding of constructs, the construct is assumed to cause variations in its indicators (i.e., items comprising the construct). On the other hand, in a formative measured construct, indicators or observed variables cause variations in the related construct. For instance, in a construct like socio-economic status, which is realized, among other indicators, through one's educational level, the area of residence, and wealth, the construct does not bring about variation in its realizations (i.e., indicators), instead; the indicators constituting the construct cause variation in the construct. In this study, test preparation was conceptualized as a formatively measured construct since it is more plausible to think of preparation practices causing variations in the construct of test preparation. In other words, it makes less sense to think of test preparation as a construct causing variation in its indicators in the same way that typical psychological constructs like intelligence and motivation do.

Before presenting the results, a brief description of the model postulated is in order. The model consists of five latent variables, two exogenous (perceptions of test uses and Perceptions of test content) and three endogenous variables (test value, the expectation of success, and test preparation). The distinction between exogenous and endogenous variables, according to Byrne (2010), is similar to the distinction between

independent and dependent variables. Pictorially, exogenous variables are those to which no arrows are directed, and fluctuations in their values are not explained by the model (Haenlein & Kaplan, 2004). On the other hand, variations in endogenous variables are explained either directly or indirectly by other variables in the model. In the model postulated in the current study (see Figures 1 & 2), variations in PTU and PTC constructs are not explained by the model as they are exogenous variables. In contrast, the three constructs of test value, expectations of success, and test preparation are modeled as endogenous variables whose fluctuations are supposed to be explained by various paths in the model. Another terminological distinction made in SEM in general and PLS-SEM in specific is between measurement (outer) models and structural (inner) models. The former refers to constructs and their corresponding indicators whereas structural models refer to the relationships between latent constructs in the model.

To perform the model evaluation in PLS-SEM, measurement models and structural models are evaluated respectively. Evaluation of measurement models is in fact about examining the reliability, validity, and factor loadings of indicators related to each measurement model. To evaluate structural models in PLS-SEM, Coefficient of determination (R^2), Predictive relevance (Q^2), Size and significance of path coefficients, f^2 effect sizes, and q^2 effect sizes are examined (J. F. Hair et al., 2016). The structural model was evaluated in another study (under review). The present study is limited to examining the structural invariance of the postulated model across two levels of the categorical, moderator variable of test takers' institutional membership. In the literature on PLS-SEM, this is discussed under heterogeneity of data.

Results

In this section, we briefly report the results of the measurement. Since the focus of this study is on the structural invariance of the

structural model, the rest of this section is given to the results of structural invariance of washback across institutions.

According to Chin (2010) measurement model evaluation in PLS-SEM is essentially about evaluating the reliability and validity of the constructs as measured in a study. To this aim, composite reliability, average variance explained, convergent and discriminant validity should be examined. In this study, the composite reliability coefficients for the constructs in the model were all above the acceptable level of .7. Convergent validity was examined through AVEs (average variance explained) and the following values were obtained: PTU (.73), PTC (.615), ES (.809), TV (.817), and Test preparation (.197). As such except for test preparation, all AVE values were close or above the required .5 threshold (see Ravand & Baghayi, 2016). The low AVE value of test preparation has to do with the fact that it was identified in the model as a formative measure. According to Hair et al. (2016), "the internal consistency perspective that underlies reflective measurement model evaluation cannot be applied to formative models since formative measures do not necessarily covary" (p.118). Furthermore, discriminant validity was assessed using Fornell-Larcker criteria, and no traitor construct was identified as all AVE values on the diagonal were higher than the values below the diagonal; which attest to the discriminant validity of the constructs (Garson, 2016). Figure 1 illustrates the outcome of both the measurement and structural models. As can be seen, all indicators have high to moderate loadings on their constructs.

As to the evaluation of the structural model, Chin (2010) maintains that what matters most in PLS-SEM is the variance explained in the target endogenous variables. As Figures 1 and 2 demonstrate, in both models around .4 of the variance in test preparation is explained in the model, which is considered a moderate degree of model adequacy. Figure 2 shows the PLS-SEM outcome of the model estimated with both the data of the two groups combined.

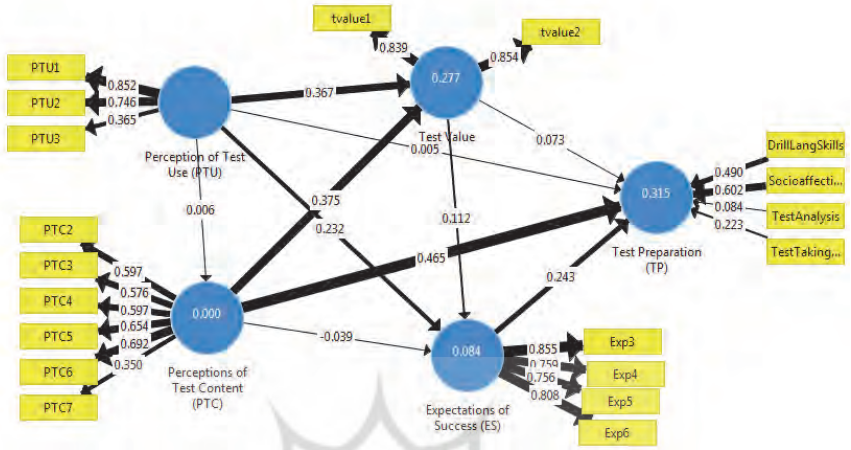

Figure 1. PLS algorithm outcome of the washback model for JDU participants

A full account of the evaluation of the structural model is not possible due to space reasons. Suffice it to say that one-third of the variation in test preparation is explained by the model, with the most reliable paths to test preparation being PTC and ES. The path coefficients from PTC and PTU are both moderate (.37 & .36), but since the path from test value to test preparation is weak, the overall mediating effect of test value is not considerable. The same is true for the mediating effect of expectation of success. In brief, when it comes to preparation for MALT, perceptions of uses and task demands are more powerful predictors of an effort than the two mediating variables of value and expectancy.

In this section, the two groups of participants are compared on each construct in the model. Secondly, results for structural model invariance across the two institutions are reported. Table 3 gives the descriptive statistics for the participants' scores on both the test perception and the test preparation scales.

Table 3.

Descriptive Statistics of SCU and JDU Groups on Constructs of Test Perceptions and Preparation

	University	*Number of items*	*Mean*	*Std. Deviation*	*Std. Error Mean*
Test Analysis	SCU	13	40.38	8.70	1.24
	JDU		38.48	11.12	1.82
Drilling Language Skills	SCU	12	33.72	7.07	1.01
	JDU		32.17	7.60	1.25
Socio-affective strategies	SCU	8	26.06	5.01	.71
	JDU		25.59	6.53	1.07
Test-taking skills	SCU	13	42.54	7.71	1.10
	JDU		43.51	9.80	1.61
Perceptions of test use	SCU	3	11.19	2.06	.29
	JDU		10.75	2.57	.42
Perceptions of test content	SCU	6	18.88	4.05	.57
	JDU		19.77	4.16	.68
Expectations of success	SCU	4	22.55	3.72	.53
	JDU		21.96	3.60	.59
Test value	SCU	2	7.08	1.61	.23
	JDU		7.01	1.81	.29

An eyeballing of the table tells that participants across the two institutions appear to be similar in both their preparation practices and their perceptions of MALT. However, to know whether washback from MALT to participants from the two institutions varies significantly, we need to use group comparison statistics. Table 4 contains the results from independent samples t-tests from between SCU and JDU participants on all the eight subscales of the two questionnaires.

Table 4.

Independent Samples T-Tests between SCU and JDU Participants

	F	*Sig.*	*t*	*df*	*Sig. (2-tailed)*	*Mean Difference*	*Std. Error Difference*
Test Analysis	5.63	.020	.88	84	.37	1.90	2.13
Drilling Language Skills	.16	.68	.97	84	.33	1.54	1.59
Socio-affective strategies	1.75	.18	.38	84	.70	.47	1.24

	F	*Sig.*	*t*	*df*	*Sig. (2-tailed)*	*Mean Difference*	*Std. Error Difference*
Test taking skills	3.16	.07	.51	84	.60	-.97	1.88
Perceptions of test use	4.76	.03	.86	84	.389	.43	.50
Perceptions of test content	.12	.72	1.0	84	.31	.89	.89
Expectations of success	.244	.62	.73	84	.46	.58	.80
Test value	.514	.47	.18	84	.85	.06	.37

We can see from Table 4 that all the t-tests are insignificant, indicating that SCU and JDU test takers do not differ significantly in their preparation for MALT. In other words, MALT washback appears to be similar across institutions. According to Table 4, SCU and DJU test takers do not have significantly different perceptions of the test content or uses, implying that test takers across institutions take MALT for similar purposes and they have similar understandings of the knowledge and skills that are crucial to success on MALT.

Before examining structural invariance, the measurement invariance of the measurement models should be established (Garson, 2016; Hair et al. 2016) because existing intergroup differences in the measurement models would invalidate conclusions arrived at regarding structural invariance. In the context of PLS-SEM, measurement invariance is performed using permutation algorithm (Garson, 2016), which compares the loadings of indicators across the levels of the moderator variables in whose possible effect we are interested. Table 5 illustrates the outcome of permutation algorithm.

Table 5.

Permutation Algorithm Output

	Outer Loadings Original (ScuJdu Uni (1.0))	Outer Loadings Original (ScuJduUni (2.0))	Outer Loadings Original Difference	Outer Loadings Permutation Mean Difference	Permutation p-Values
DrillLangSkills -> Test Preparation (TP)	0.812	0.440	0.372	0.056	0.343
Exp3 <- Expectations of Success (ES)	0.894	0.819	0.074	0.005	0.411
Exp4 <- Expectations of Success (ES)	0.664	0.868	-0.204	0.004	0.192
Exp5 <- Expectations of Success (ES)	0.747	0.717	0.030	0.009	0.808
Exp6 <- Expectations of Success (ES)	0.891	0.673	0.219	-0.001	0.071
PTC2 <- Perceptions of Test Content (PTC)	0.501	0.664	-0.162	0.015	0.567
PTC3 <- Perceptions of Test Content (PTC)	0.775	0.318	0.457	0.000	0.069
PTC4 <- Perceptions of Test Content (PTC)	0.464	0.712	-0.248	0.001	0.382
PTC5 <- Perceptions of Test Content (PTC)	0.812	0.472	0.340	0.020	0.109
PTC6 <- Perceptions of Test Content (PTC)	0.458	0.861	-0.403	0.014	0.066
PTC7 <- Perceptions of Test Content (PTC)	0.465	0.036	0.428	-0.002	0.191
PTU1 <- Perception of Test Use (PTU)_	0.748	0.759	-0.011	0.048	0.932
PTU2 <- Perception of Test Use (PTU)_	0.742	0.737	0.005	0.033	0.989
PTU3 <- Perception of Test Use (PTU)_	-0.062	0.801	-0.863	-0.029	0.131
Socioaffective -> Test Preparation (TP)	0.700	0.784	-0.084	0.070	0.749
TestAnalysis -> Test	0.372	0.643	-0.271	0.020	0.512

	Outer Loadings Original (ScuJdu Uni (1.0))	Outer Loadings Original (ScuJduUni (2.0))	Outer Loadings Original Difference	Outer Loadings Permutation Mean Difference	Permutation p-Values
Preparation (TP)					
TestTakingSkills -> Test Preparation (TP)	0.463	0.790	-0.327	0.040	0.495
tvalue1 <- Test Value	0.837	0.801	0.036	0.008	0.797
tvalue2 <- Test Value	0.845	0.897	-0.053	0.003	0.666

The far left-hand column lists the indicators along with their corresponding construct, and the far right column gives the permutation p-values, which tests the significance of the difference between the loading of each item on its pertinent construct across JDU and SCU participants. As can be seen in the column, none of the p-values is significant. This is evidence that measurement equivalence is ensured across the levels of the moderator variable of the institution.

Lack of difference between the two groups of test takers on individual subscales, however, does not readily mean that a model of washback would hold the same across the two samples of test takers, mainly because of possible, complex interactions between and among various endogenous and exogenous variables in the model. To know the characteristics of a theoretical model of washback across the two samples of participants, model estimation was done separately for the two groups of test takers.

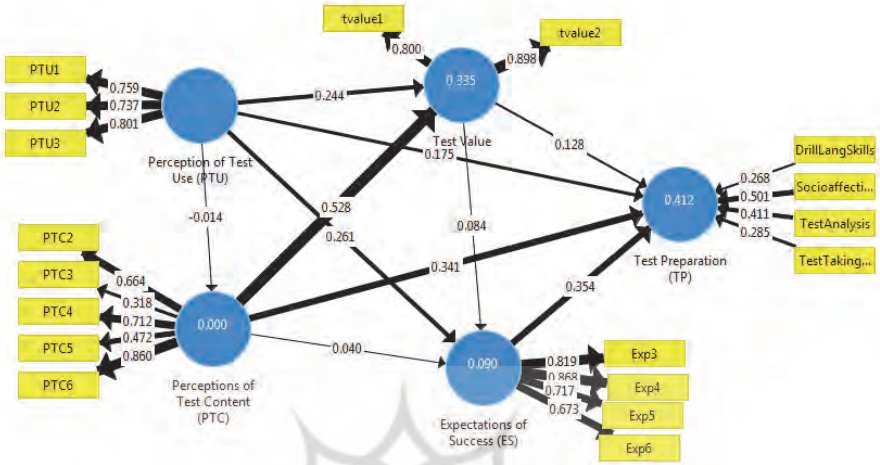

Figure 2. PLS algorithm outcome of the washback model for JDU participants

Figure 2 displays the PLS-SEM output for JDU participants. Values in the blue balls are r^2 values, and those on the paths are path coefficients or correlations between constructs in the model. The thickness of the arrows is commensurate with the strength of path coefficients. As can be seen in Figure 2, taken together 41 percent of variance in the target endogenous variable, test preparation, is explained via the postulated model and the most significant path coefficient is between perceptions of test content to test value (r= .528), indicating that knowledge or awareness of test demands is associated with more value placed on taking MALT. Also, the two constructs predicting the most substantial portion of the variance in preparation for MALT are expectations of success and perceptions of test content, respectively.

Figure 2 demonstrates the output from PLS-SEM analysis for participants from SCU. As noted earlier, values in blue balls are r^2 values, and those on the paths are path coefficients or the correlations between constructs.

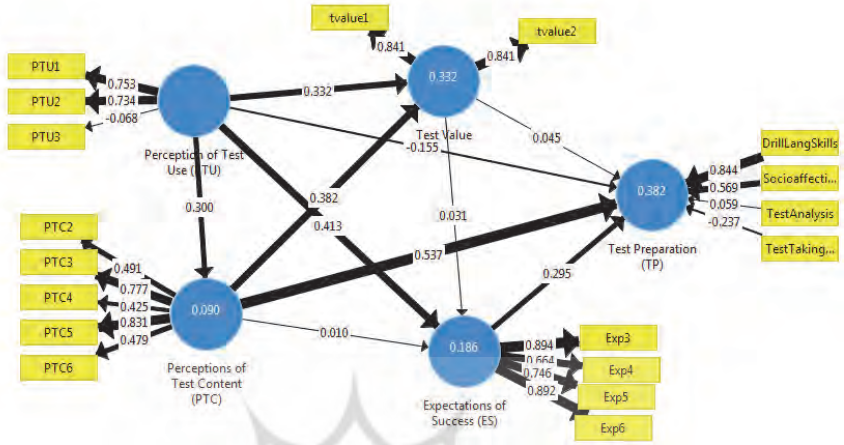

Figure 3. PLS algorithm outcome of the washback model for SCU participants

According to Figure 3, the overall variance in test takers' preparation practices explained by the model is .38, which is a medium R square value (J. F. Hair et al., 2016). The most robust paths are those from *perceptions of test content* to *test preparation* (r= .537), *perceptions of test uses* to *expectations of success* (r=.413), *perceptions of test content* to *test value* (r=.382), *perceptions of test uses* to *test value* (r= .33), and *expectations of success* to *test preparation* (r=.295), respectively. PTU and *test value* do not seem to make a considerable direct contribution to explaining variation in *test preparation*.

To know whether each construct has made a significant unique contribution to the model, we need to assess f^2 effect sizes. According to J. F. Hair et al. (2016), the f2 effect size in the context of PLS-SEM tells us "how much a predictor construct contributes to the R^2 value of a target construct in the structural model" (p.198). Table 5 contains the f2 effect size values via the bootstrapping procedure, which is a statistical resampling procedure (Westfall & Young, 1993) of the family of robust statistics that provide the power of parametric statistics without necessitating the stringent requirements of parametric statistics (Larson-

Hall & Herrington, 2010). Especially, bootstrapping is "a nonparametric procedure that randomly resamples from an observed data set to produce a simulated but more stable and statistically accurate outcome" (Plonsky, Egbert & Laflair, 2014, p. 1). In other words, "the observed data are repeatedly used, in computer-intensive simulation analysis, to provide inferences. In simple terms, resampling does with a computer what the experimenter would do in practice if it were possible: he or she would repeat the experiment" (Westfall & Young, 1993, cited in Larsen-Hall & Herrington, 2010, p. 379). Statistical analyses based on bootstraps have been shown to be notably more potent than conventional parametric statistics in reducing the probability of making Type II errors in quantitative research (Larson-Hall & Herrington, 2010; Plonsky, Egbert & Laflair, 2014).

To conduct multi-group analysis in PLS-SEM, several approaches have been proposed, usually discussed "modeling heterogonous data" (Hair et al., 2016, p. 243). "Failure to consider heterogeneity can be a threat to the validity of PLS-SEM results since it can lead to incorrect conclusions" (Hair et al., 2016, p. 244). Data heterogeneity is considered either unobserved or observed. Unobserved heterogeneity refers to the cases "when the true sources of heterogeneity in data sets are unknown" (ibid, p. 244). Finite mixture PLS (FIMIX-PLS) is utilized to model unobserved heterogeneity. Observed heterogeneity is of concern when researchers want to explore the effect of a categorical moderator variable on the path coefficients between exogenous and endogenous variables of the model. The multi-group analysis in the present study is considered a case of observed heterogeneity since we sought to know if the categorical variable of the institution, with two levels (i.e., SCU & JDU), moderates the relationships in the model.

To conduct multi-group analysis with categorical, moderator variables, parametric and non-parametric approaches have been proposed. The permutation-based approach (Chin, 2003) and the approach proposed by Henseler (2007, cited in Sarstedt, Henseler, & Ringle, 2011) are non-parametric and the approach put forward by Keil

et al. (2000) is parametric. SmartPLS version 3 conducts both parametric and non-parametric analyses. Tables 5 and six below illustrate the outcome of non-parametric and parametric comparisons, respectively.

Table 5.

Bootstrapping Results: f2 Effect Sizes for Both Groups

	Path Coefficients (SCU)	*Path Coefficients JDU*	*t-Values (SCU)*	*t-Values (JDU)*	*p-Values (SCU)*	*p-Values (JDU)*
Expectations of Success (ES) -> Test Preparation (TP)	0.295	0.354	1.028	1.326	0.304	0.185
Perception of Test Use (PTU)_ -> Expectations of Success (ES)	0.413	0.261	1.498	1.301	0.135	0.194
Perception of Test Use (PTU)_ -> Perceptions of Test Content (PTC)	0.300	-0.014	1.156	0.052	0.248	0.959
Perception of Test Use (PTU)_ -> Test Preparation (TP)	-0.155	0.175	0.673	0.653	0.501	0.514
Perception of Test Use (PTU)_ -> Test Value	0.332	0.244	1.737	0.916	0.083	0.360
Perceptions of Test Content (PTC) -> Expectations of Success (ES)	0.010	0.040	0.047	0.113	0.962	0.910
Perceptions of Test Content (PTC) -> Test Preparation (TP)	0.537	0.341	2.048	0.948	**0.041**	0.344
Perceptions of Test Content (PTC) -> Test Value	0.382	0.528	2.589	2.398	**0.010**	**0.017**
Test Value -> Expectations of Success (ES)	0.031	0.084	0.151	0.204	0.880	0.839
Test Value -> Test Preparation (TP)	0.045	0.128	0.227	0.373	0.821	0.709

The f2 effect sizes show that for the SCU group, perceptions of test content (PTC) construct make a significant unique contribution to test

preparation (p= .04). Similarly, the f^2 effect size for the path from PTC to test value is significant (p= .01) for the SCU group. For the JDU participants, the only significant f2 effect size is from perceptions of test content to test value.

Juxtaposing the two models, there appear to be two noticeable differences. First, for JDU students (Figure 2), the most reliable path coefficient is between perceptions of test content and test value, whereas for SCU students (Figure 3), the most significant correlation is between perceptions of test content and test preparation. Moreover, the path coefficient from test value to test preparation for JDU group is .128, while it is .045 for the SCU participants. To see whether the observed differences between the two models are significant, we carried out Multi-Group Analysis (MGA) in PLS-SEM. To do so, the bootstrapping procedure was used. This procedure compares the observed difference between the two models with the universe of differences between numerous pairs (500) of similar groups.

Table 6.

Parametric Tests for the Differences in the Path Coefficients for the Two Models

	Path Coefficients-difference (SCU-JDU	*t-Value (SCU vs. JDU)*	*p-value (SCU vs. JDU)*
Expectations of Success (ES) -> Test Preparation (TP)	0.059	0.146	0.884
Perception of Test Use (PTU)_ -> Expectations of Success (ES)	0.153	0.425	0.672
Perception of Test Use (PTU)_ -> Perceptions of Test Content (PTC)	0.314	0.827	0.411
Perception of Test Use (PTU)_ -> Test Preparation (TP)	0.330	0.947	0.347
Perception of Test Use (PTU)_ -> Test Value	0.088	0.279	0.781
Perceptions of Test Content (PTC) -> Expectations of Success (ES)	0.029	0.075	0.940
Perceptions of Test Content (PTC) -> Test Preparation (TP)	0.195	0.455	0.650

	Path Coefficients-difference (SCU-JDU	*t-Value (SCU vs. JDU)*	*p-value (SCU vs. JDU)*
Perceptions of Test Content (PTC) -> Test Value	0.147	0.580	0.563
Test Value -> Expectations of Success (ES)	0.053	0.125	0.900
Test Value -> Test Preparation (TP)	0.083	0.225	0.823

In a nutshell, parametric tests for the differences in the path coefficients for the two models, reported in Table 6, indicate that path strength differences between the two models are not significant, indicating that the proposed model of washback explains a moderately high level of variation in test preparation across institutions. Some scholars place less trust in the outcomes of parametric approaches to Multi-group comparisons because they are incompatible with the underlying logic of PLS-SEM (Sarstedt, Henseler, & Ringle, 2011).

Discussion and Conclusions

This study investigated the often taken for granted idea that there is an interaction effect between the quality of an educational institution where one studies and washback from high stakes tests (see Xie & Andrews, 2013; Watanabe, 2004). In other words, we put to an empirical test the idea that high stakes tests induce more intense washback to students from lower-tier universities than to those from top-tier universities, as maintained by Xie and Andrews.

Regarding the first research question, which was about the adequacy of perceptions and value in predicting test preparation, results from PLS-SEM analysis revealed that the postulated washback model explains a moderate amount of variation in test takers' preparation for MALT, which indicates that test takers' reasons for taking tests combined with their perceptions of test content and demands can partly explain their test preparation behavior. This finding is following Xie and Andrew's (2013) findings. Xie and Andrews found that test takers endorsing high stakes

testing engaged in more intense test preparation. The finding is also consistent with Alderson and Wall's (1993), who foresaw the relevance of motivation in predicting test washback.

Furthermore, it was found that the path from test perceptions to test value is significant for JDU participants but not for those from the SCU. This might be since students in SCU have uniform access to MALT sources both through the departmental physical and electronic resources as well as through direct access to graduate and postgraduate students who have already passed MALT. As such, for SCU students, it is not the knowledge of test demands that determines the value of test taking. On the other hand, for JDU test takers, who are less exposed to information MALT in their immediate educational environment, knowledge of MALT content and demands accrues only to those who seek it; those who value taking MALT. This might be counted as evidence in support of the proposition that washback effect of testing takers' perceptions of tests is moderated by the educational environments, though one has to be cautious in making generalizations based on their rather slim evidence.

Moreover, it was found that the constructs in the proposed washback model explained slightly more of the variance in test preparation for the JDU participants than the one for SCU test takers (see Figures 3 and 3). This is also aligned with past research (Xie & Andrews, 2013; Hamp-Lyons, 1998; Wall, 1996). The literature has it that when a test is perceived to be within the proximal zone of difficulty and challenge, it is more likely to induce washback (Hamp-Lyons, 1998; Wall, 1996). With the same logic, it might be plausible to think that MALT fits better with the zone of the proximal challenge of students from low tier universities.

Concerning the overall structural invariance of the washback model, we did not find substantial evidence in support of the moderating effect of the educational environments in the intensity of test washback. It might be that universities with different ranks should not necessarily be deemed considerably different regarding their educational environments. Another possibility is that in the era of social media and communication, the spread of information might have rendered boundaries between

institutions fluid. As such, the virtual world might have a more powerful influence on learners' perceptions than the physical, institutional setting. This is a serious possibility given that English major students are relatively highly digitally literate because of the opportunities afforded to them via access to English, which might have leveled the ground for test takers across various institutions by creating equal or similar access to information, including information about the test. Still another possibility might have to do with the very physical proximity of the two universities and their being based within the same socio-cultural milieu of the same city. Finally, it might be that the washback of MALT is so intense that it overrides or neutralizes variations in test takers' institutional backgrounds.

Washback is known to be a highly contextual phenomenon, dependent on social, cultural, individual and test factors (Watanabe, 2004a). Perhaps, these contextual factors are responsible for the discrepancy found between the findings of the present study and what is assumed to be the case in the literature (Hamp-Lyons, 1998; Xie & Andrews, 2013). In particular, the function that a test serves is vital in the kind of washback that might be produced. For instance, in Xie and Andrews' study, it was taken for granted that low test takers from low tier universities are more prone to test washback from CET, an exit test, than their counterparts in top tier ones. MALT, the test under consideration in this study, however, is an admission test. In an exit test, all test takers are required to take the test to graduate, whereas taking an admission test is voluntary. Hence, possibly those who choose to take an admission test, MALT, in this case, share motivational characteristics that offset differences stemming from universities where they are graduating from. That said, whether and how exit and admission test produce differential washback to test takers' perceptions and learning practices is open to further inquiry.

Implications, Limitations and Further Inquiry

The present study might hold some implications for policymakers, test designers, and teachers, especially those involved in preparing test takers for high stakes tests. At the level of policy, it is common practice for policymakers to adopt a one-size-fits-all approach in formulating and implementing language testing policies, without due attention to the particulars of various educational institutions.

Previous research has shown that test takers' perceptions of tests are related to testing validity (Qin, 2011). Although the evidence was not entirely conclusive, there were some indications in the present study that for test takers from various academic institutions, perceptions of tests, and their subsequent test preparation practices, might not be substantially uniform. The combination of these two premises would boil down to the following. If test-takers perceptions are related to validity and if such perceptions are affected by institutional climates, then institutional environments can indirectly bear on the validity of tests. This might come across as rather bizarre from a pure psychometric vantage point but viewing validity from a Messickian perspective, where social values and consequences are to be seen as aspects of the validity of test-based inferences (Messick, 1989), it makes sense.

Test preparation is often seen as illegitimate, unethical, and harmful to the final missions of education (Crocker, 2005; Gebril & Eid, 2017; Hamp-Lyons, 1998), yet; it is prevalent. It is estimated that nearly half of money families spend on education goes to test preparation (Gebril & Eid, 2017), indicating that it is highly prevalent. Therefore, denying the existence of test preparation industry or questioning its effectiveness does not get us anywhere. Preferably, for teachers to harness the vast potential of the test preparation industry in the service of educational objectives, they need to be aware of how test takers' motivational differences and their different understanding of test content and demands are related to their test preparation practices. To do so, teachers in test preparation courses should also be mindful of how school climate and culture might shape test takers' attitudes and perceptions of the tests. Research shows

that test takers' perceptions of test fairness are both "culturally and contextually situated" (Jang, 1991, p. 3). If contextual and institutional forces lead test takers to perceive of a test as unfair, it would have far-reaching consequences for the value they attach to the test results as well as for their resilience in test directed language learning.

As with all research, this study had its share of limitations. For one thing, due to logistic problems, we sampled participants only from two universities in the same province. It is not conceptually implausible to think that had the two universities differed not only concerning their ranking but also in their sociocultural environments, and the findings might have been otherwise. This possibility is consistent with the highly contextually bound nature of washback (Watanabe, 2004).

Typical of most L2 research (LaFlair, Egbert, & Plonsky, 2016) the other limitation of this study was the sample size. In particular, the JDU sample was somehow below the minimum sample size required. Although thanks to its powerful bootstrapping mechanism, PLS-SEM is said to be resilient in the face of small sample sizes (Hair et al., 2016), studies with larger sample sizes would likely provide us with richer insights and more generalizable patterns. Furthermore, studies adopting mixed methods research designs may produce a more nuanced understanding of how institutional climates bear on test washback and preparation. Finally, though steps were taken to maximize the validity of the instruments used in the study, we believe that with more valid instruments, which can capture the nuances of both institutional contexts and test takers' perceptions and preparation, future inquiry can uncover more hidden dimensions of the complex interactions between tests, test takers, and educational environments.

References

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

Byrne, M. B. (2010). *Structural equation modeling with AMOS: Basic concepts, applications, and programming*: London: Lawrence Erlbaum Association–Publisher.

Chin, w. (2010). How to write up and report PLS analyses. In E. Vinzi, W. Chin, J. Henseler, & H. Wang (eds). *Handbook of partial least squares* (pp. 655-690). *Springer*

Cheng, L., Sun, Y., & Ma, J. (2015). Review of washback research literature within Kane's argument-based validation framework. *Language Teaching, 48*(4), 436-470.

David, A. (2016). Investigating washback to the learner from the IELTS test in the Japanese tertiary context. *Language Testing in Asia, 6,* 2-20. doi:10.1186/s40468-016-0030-z

Entwistle, N. J. (1991). Approaches to learning and perceptions of the learning environment. *Higher education, 22*(3), 201-204.

Farhady, H., & Hedayati, H. (2009). Language assessment policy in Iran. *Annual Review of Applied Linguistics, 29,* 132-141.

Freiberg, H. J. (1999). *School climate: Measuring, improving, and sustaining healthy learning environments*: Psychology Press.

Haenlein, M., & Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding statistics, 3*(4), 283-297.

Hair, J., Sarstedt, M., Hopkins, L., & Kuppelwieser, V. (2014). Partial least squares structural equation modeling (PLS-SEM) An emerging tool in business research. *European Business Review, 26*(2), 106-121.

Hair, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*: Sage Publications.

Hamp-Lyons, L. (1998). Ethical test preparation practice: The case of the TOEFL. *TESOL Quarterly, 32*(2), 329-337.

Kaur, A., Noman, M., & Awang-Hashim, R. (2017). The role of goal orientations in students' perceptions of classroom assessment in higher education. *Assessment & Evaluation in Higher Education, 43*(3), 1-12.

Larson-Hall, J., & Herrington, R. (2010). Examining the difference that robust statistics can make to studies in language acquisition. *Applied Linguistics, 31*(3), 368-390.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement*. New York: Macmillan.

Plonsky, L., Egbert, J., & Laflair, G. T. (2014). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics, 36*(5), 591-610.

Qin, X. (2011). Is Test Taker Perception of Assessment Related to Construct Validity. *International Journal of Testing, 11*(4), 324-348. doi:10.1080/15305058.2011.589018

Ravand, H., & Baghaei, P. (2016). Partial least squares structural equation modeling with R. *Practical Assessment, Research & Evaluation, 21*(11), 1-16.

Sarstedt, M., Henseler, J., & Ringle, C. M. (2011). Multigroup analysis in partial least squares (PLS) path modeling: Alternative methods and empirical results. In *Measurement and research methods in international marketing* (pp. 195-218). Emerald Group Publishing Limited.

Shin, S.-K. (2005). Did they take the same test? Examinee language proficiency and the structure of language tests. *Language Testing, 22*(1), 31-57.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: washback effect over time. *Language Testing, 13*. doi:10.1177/026553229601300305

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & McPhail, R. (2013). Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'. *Assessment & Evaluation in Higher Education, 38*(1), 44-60.

Spratt, M. (2005). Washback and the classroom: the implications for teaching and learning of studies of washback from exams. *Language Teaching Research, 9*(1), 5-29.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing, 10*(1), 41-69

Watanabe, Y. (2004a). Methodology in washback studies *Washback in language testing: Research contexts and methods* (pp. 19-36). In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods (pp. 19-36)*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Watanabe, Y. (2004b). Teacher factors mediating washback. In L. Cheng, Y. Watanabe, & A. Curtis (Eds.), *Washback in language testing: Research contexts and methods (pp.* 129-146). Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Westfall, P. H., Young, S. S., & Wright, S. P. (1993). On adjusting P-values for multiplicity. *Biometrics, 49*(3), 941-945.

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary educational psychology, 25*(1), 68-81.

Xie, Q., & Andrews, S. (2013). Do test design and uses influence test preparation? Testing a model of washback with structural equation modeling. *Language Testing, 30*(1), 49-70

Zhan, Y., & Andrews, S. (2014). Washback effects from a high-stakes examination on out-of-class English learning: insights from possible self-theories. *Assessment in Education: Principles, Policy & Practice, 21*(1), 71-89.