Language Testing in Asia

RESEARCH Open Access

A structural equation modeling of English tests' social and educational consequences: exploring target, leverage, risk, and critical variables

Mahmood Khosravani[1], Morteza Rostamian[1,2*], Hamid Ashraf[1] and Khalil Motallebzade[1]

*Correspondence: m.rostamian.edu@gmail.com

[1] Department of English, Torbat-e Heydarieh Branch, Islamic Azad University, Torbat-e Heydarieh, Iran
[2] English Department, Faculty of Medicine, Social Development and Health Promotion Research Center, Gonabad University of Medical Sciences, Gonabad, Iran

Abstract

Though the emergence of standardized testing promoted classical approaches such as summative achievement tests, it brought about serious criticisms regarding the stakes of these tests and their consequences on different dimensions of education. Exploring these consequences attracted researchers' attention in education. A mass of studies are conducted, but due to the dynamic and intricate nature of the consequences, they suffer from some severe deficiencies, mainly in methodology and epistemology. Thus, the current research adopted an innovative approach to employ structural equation modeling in SPSS and matrix of crossed impact multiplications in MicMac to explore factors of test consequences of English module of Iranian universities entrance exam. Four groups of participants were selected through convenience sampling from different Iranian stakeholders for scale development, reliability estimation, exploratory factor analysis (EFA), and crossed impacts analysis, respectively. Reliability calculation was carried out ($r = 0.89$), and assumption of EFA for factorability ($KMO = 0.796$ and $p = 0.001$) was met. EFA was run, and 19 variables (factor loading ≥ 0.60) in three major factors were identified. Also, crossed impact analysis showed that these factors fall in an unstable system of effects, inside input factors (critical and environmental), bidimensional variables (risk and target), output variables, and leverage variables. Integrating structural equation modeling and structural interpretations resulted in development of a validated model and uncovered the nature of the variables within it. These information can feed subsequent interpretations, including policy makings and future studies for educational and management purposes. The findings have educational and statistical implications for different stakeholders.

Keywords: Factor analysis, Matrix of crossed impacts, Structural equation modeling, Test consequences

Introduction

The emergence of standardized testing by the end of the 1970s and its high power in teaching and learning transformations brought about serious criticism due to the stakes of the tests and several intended and unintended severe consequences, including

narrowed curriculum, teaching to the test, learning for the test, and a wide array of social consequences (Chen et al., 2020). According to Green (2020), this view that test consequences are essential is long discussed in language literature, where terms such as washback, impact, and consequences were coined to explain a broad range of educational and social consequences that may result from test use. Consequences of language tests are the result of decisions that are made based on test scores and their interpretation, because they exert influence on tastes' future channel of instruction, university graduation, certification for employment, and migration (Chapelle, 2020). Messick (1989) was the first one that considered the social functions that tests serve as one angle of look for investigating the utility of a test and its social consequences. For the first time, Messick (1989) examined social values and their consequences at the center of any validity theory by emphasizing the importance of studying consequences to understanding the meaning of test scores. Later, researchers and scholars in the field, such as Moss et al. (2005), raised this awareness in the discipline that scores from a particular test can be interpreted and used in multiple ways, depending on individual responses and the contexts of assessment. Since then, many scientists have extended research on consequential validity among which is Kim (2017), which blended different theories in the discipline to arrive at a comprehensive framework for assessing the consequential validity of tests which encompasses educational consequences, meaningfulness, directness, transparency, fairness, and usability. According to Bachman and Palmer (1996), under the umbrella term of consequential validity of the large-scale and high-stakes test, several phenomena affect the interests of stakeholders and those directly related to these tests, which entail two significant category of consequences: washback and impact. These terms (washback, impact, and consequences) are used in different research fields and encompass different dimensions of the research in the testing discipline. However, Kiani et al. (2009), illuminated by Bachman and Palmer's taxonomy (Bachman & Palmer, 1996), stated that the consequences of tests on teaching and learning are viewed as washback effects, while the consequences on individual stakeholders such as learners, teachers, parents, test takers' family members, and society are considered as impact. Thus, considering that impact and washback are established disciplines of research in language testing (Cheng et al., 2015). The term "consequences" is used as a general term under which impact and washback are discussed in our research.

Like other high-stakes tests, English modules of the Iranian universities entrance exam (IUEE) exert consequences of different kinds on different peoples and educational components involved. In spite of these consequences, a significant gap in the literature exists which encompasses the following issues: (a) almost all researches conducted in EFL contexts and Iran have adopted a one-size-fits-all approach toward their investigation by either employing a global scale of test consequences or a global methodology in different stages of research (Farokhipour et al., 2020) while asserting the transition from method to post-method, Kumaravadivelu (2005) states that adopting a one-size-fits-all approach ignores local knowledge and context-specific problems of learning, learners, and society; (b) adopted tools and instruments, even those conducted in ESL contexts, are faced with construct underrepresentation to ignore test effects out of teaching and learning realm such as social or emotional consequences (Alavi & Masjedlou, 2017), (c) most of the researches conducted in the

filed fail to employ a credible methodology for their component extraction or validation; and (d) last and most importantly, controlling diverse effect of any phenomenon calls for not only discovering and validating factors involved but also the nature of the variables to be eliminated, regulated, strengthened, or modified. Besides, Kuang (2020) found that test consequences are a complicated phenomenon that cannot be captured by adopting a single and simple approach in research. Jamalifar et al. (2021) also echoed similar findings and concluded that test effects cannot be deeply grasped by only qualitative or quantitative measures and needs integration of methods.

Accordingly, the current research adopted an innovative methodology to employ a context-embedded approach toward investigating test consequences in the Iranian context. According to Mahmudi (2014) (Mahmud: The washback effect of Iranian national university entrance exam (INUEE) on pre-university English teaching and learning, Unpublished doctoral dissertation), the studies carried out on the phenomenology of test effects and influences fail to utilize a manifest theoretical basis for their view of test validity and test effect. In other words, the present study makes significant contributions to research on test effects considering previous washback studies in language assessment have been conducted independently of validity (Bachman, 2005; Cheng, 2008). However, this study fills this research gap by connecting test effect evidence to those from validation framework by adopting the theoretical bases of Messick's theory of validity (discussed below) and utilizing an argument-based view toward inclusion and/or exclusion of factors and variables in the construct of validity. Besides, the current research investigates the consequences of the IUEE by including perspectives of test developers, test users, and educational managers to entail all stakeholders. Among other novelties of the present study is the inclusion of both washback and impact effects. On top of that, in all model-making studies through structural equation modeling, the development of the graph of the model is the final stage of the research. However, the current study has employed a multiphase, multi-method approach and utilized a matrix of crossed impact effects to uncover the nature of the variables in the model and the way they interact with other variables and factors in the model. Structural interpretation of a developed and validated model sheds light on the nature and behavior of the variables and feeds any subsequent interpretation and decision-making. The findings of this stage categorize variables into different practical types which help practitioners and decision-makers to use appropriate ones, aimed at manipulating the system. To this purpose, structural equation modeling (exploratory factor analysis) was used as a significant component extraction method. This helps in eliminating some of the previous methodological deficiencies, because this method of modeling not only extracts the factors and items involved in test consequences but also validates them. Furthermore, a matrix of crossed impact multiplications was used to shed light on the nature of the identified variables, which lays the groundwork for scrutinizing and explaining the current system of effects and provides the stakeholder with an outlook to predict the system by manipulating the factors and variables. Thus, the following research questions are formulated:

1) What are the main factors of test consequences in the English module of IUEE?

2) What are strategic variables in the model of the test consequences and how they interact with the Iranian testing context?

Literature review

General background

The history of inquiring into test consequences dates back to the 1950s and 1960s when practitioners and scholars in general education started asserting that high-stakes tests may influence teachers, learners, and other stakeholders. For instance, Vernon (1956) reported that teachers show their inclination toward ignoring the content that enjoys little chance of appearing in the exam through distorting and changing the curriculum. A reverse effect was also reported by Davies (1968) that emphasized the effects of the test on teaching and curriculum by discussing that tests and test materials are likely to be used as teaching devices to narrow down the teaching content. It was about two decades later when a similar phenomenon received attention in language learning and teaching disciplines. Delving into the concept of washback, for instance, Swain (1985) stated that language teachers would teach content to their students in consistence with the test. Besides, Bailey (2018) reporting a context-specific study in Japan stated that the Japanese university entrance exam creates a strong impact on teachers and students by saying that the influence made them tailor their classroom activities to the demands of the test. Furthermore, Hughes (1989) asserted that awareness of teachers, as test developers, of the effect of testing on learning outcomes is a critical issue to be considered. Since then, the concepts of consequential validity and its construct became established, and then, the empirical investigation of these consequences, and their related terms and concept became popular.

Theoretical background

From a more theoretical perspective, according to McNamara (2006), Messick's (1989) validity model was the most influential in language assessment research, especially in validation studies. From validity perspective, the impact is a quality that needs consideration in test design. In addition, Bachman and Palmer (1996) put "impact" under the notion of test usefulness to sit beside qualities such as reliability, construct validity, authenticity, and interactiveness, arguing that the overall usefulness of a test is a function of these qualities. Thus, a test developer must design them into the test and prioritize them based on the practicality of the testing situation. The pitfall inherent in this conceptualization was that the way these qualities interact was unknown in the model. Later, Bachman and Palmer (2010) proposed an assessment-use-argument framework and linked consequences of test use to construct validity. This investigation and even later ones such as Doe (2013) (Doe: Validating the Canadian academic English language assessment for diagnostic purposes from three perspectives: scoring, teaching and learning, Unpublished doctoral dissertation) focused primarily on the decisions related to scoring, teaching, and learning. Therefore, unintended social consequences and ethical considerations were not included. However, Cizek (2012) claimed that unintended social adverse consequences should be incorporated into a code of ethics or fair testing

practice in the issue already established by Bachman (2005). Additionally, Alderson and Wall (1993) were among the earliest washback researchers, but in the same year, Hughes (1993) (Hughes: Backwash and TOEFL 2000, Unpublished manuscript) proposed a trilateral model of participants, process, and products in washback studies. This model noted that the nature of a test may first influence participants' perceptions and attitudes toward their teaching and learning tasks, which, in turn, affects teaching and learning processes and outcomes. Integrating views from these two studies, Bailey (1996) presented the first model of washback, depicting how a test may influence various participants engaging in different processes. Later on, Green (2007) proposed a renewed model to assert that some other factors are involved in or interact with the effect of language testing and learning, such as a test taker's perception. In the same year, Shih (2007) investigated the washback of an EFL test on English learning in Taiwan and proposed a washback model of students' learning that entails influences of various factors related to the context, the student, and the test.

Related research

From an empirical point of view, a vast bulk of studies are conducted on phenomenology and modeling of test consequences in different countries. Ali and Hamid (2020) investigated factors associated with adverse washback effects in secondary school in Bangladesh. They collected data through interviews with language teachers, official documents, and other qualitative sources such as media reports. In addition to that, a bulk of empirical data were examined to arrive at critical qualitative propositions regarding tests' adverse washback. It was found that some educational and sociopolitical factors are involved in the adverse washback effect. Failing to use a specified theoretical framework and an analytical tool or criteria was among deficiencies of this study. Findings were also presented merely as qualitative reports, and their generalization and validation were less discussed. Larsson and Olin-Scheller (2020) also examined the washback effects of the national test on the upper secondary Swedish context. Adopting Stephen Ball's theory of policy enactment as the theoretical foundation of the research and utilizing interviews as a major data collecting method, the study revealed that the exam, due to some social forces, exerts debilitative effects on some language skills while leaving others less affected. The study used qualitative reports in data collection and adopted a narrow and fractional perspective for analysis and report of findings. Similarly, Dong (2020) conducted a study on washback mechanisms of high-stakes tests and examined the relationship between learners' perception and their practices and outcomes through structural equation modeling. The study revealed that issues such as impact and validity affect various dimensions of teaching differently. This study did not cover sociopolitical and consequential factors and variables however. Previously, Hung and Huang (2019) ran a washback study about language curriculum to examine the relationships between washback and learner characteristics such as education, gender, and proficiency level. The findings showed that washback has a significant influence on students' educational (such as proficiency) and noneducational (such as self-image) characteristics.

Similar studies were carried out in the Iranian context. Tabatabaei and Safikhani (2011) studied the washback effect of a university entrance exam on EFL teachers' methodology and test development. They found that the exam impacted the EFL teachers' methods

and test development. The main criticism was that the questionnaire and other tools used in the study did not undergo any statistical validation process, save the fact that the study underlined only the educational dimension of test consequence. Mahmoudi (2015) worked on the washback effect of the Iranian National University entrance exam on pre-university English teaching and learning. She found that many factors affect the process of English language learning and teaching in pre-university schools and proposed a model of washback. However, developing a model out of mere descriptive statistics without any construct evaluation procedure may not result in a valid and credible model. To explore the washback effect of the English module of the Ph.D. entrance exam, Rezvani and Sayyadi (2016) used the interview as the primary tool of their study which limits findings and their generalizations. Furthermore, Sadighi (2018) examined the possible washback effect of the university entrance exam. The findings were consistent with the earlier studies concerning the washback impact of the high-stakes exams on the teaching materials. This study was inclined toward the washback dimension of the effect because it included educational factors while ignoring social, emotional, and possible consequences of the test. Considering the mere education-centered direction of washback studies, Estaji and Ghiasvand (2019) studied the washback effect of the IELTS examination on Iranian EFL teachers' professional identity at the expense of failing curriculum and the core associated effects. Moreover, Jamalifar et al. (2021) studied the effect of a high-stake test from a qualitative perspective.

Accordingly, the current research was an innovative attempt to embrace all dimensions of the effects through a credible and validated methodology employing structural equation modeling. Besides, the present study is the first one that went beyond mere phenomenology and explained the existing condition and adopted a matrix of crossed impact multiplications to identify the nature of the variables involved and put forward some suggestions for predicting, regulating, modifying, and controlling the test consequences.

Context of the study

Iran is an EFL context in which English is taught at primary and secondary middle schools as part of their regular curriculum. Recently, the curriculum has undergone dramatic changes through the Iranian National Curriculum, a program aimed at promoting language skills and communicative competence. However, this increased inconsistencies the curriculum had with IUEE because the exam is mainly dependent on grammar and vocabulary while ignoring language skills. This condition has resulted in an intricate phenomenon favoring diverse washback effects and negative consequences beyond the educational context. Since the number of the variables involved is very large, the previous studies have examined some parts of them. However, the current study adopted a multi-method, multiphase design to embrace other stakeholders and dimensions.

Methods

The current research adopted a multiphase, multi-method approach in data collection and data analysis. Accordingly, in the earlier phase of the research, a structural equation modeling was run, and factors and variables were extracted, validated, and modeled. In the second phase, however, a matrix of crossed impact multiplications was utilized to

explain the nature of the variables and the quality of their interaction in the system of the exam's effects.

Participants

In this study, considering twins of time and expenses limitations and requirements of the research design, five main groups of participants were selected among different stakeholders including teachers, students, educational managers, and parents, through convenience sampling from Tehran, Mashhad, and Birjand. The first group encompassed nine language experts (nine male and female assistant professors of TEFL from the Azad University in different branches to include both genders) used in the questionnaire development stage. The second group was employed in reliability estimation of the primarily generated scale of the study, where thirty-two male and female participants (twenty-four students, and eight teachers) answered the items generated in a Likert questionnaire. The third group of the participants used in the exploratory factor analysis phase encompassed one-hundred and fifty-seven male and female subjects (one-hundred and ten students, thirty-six teachers, five educational managers, and six parents). The fourth group of participants ($N = 240$) encompassed 190 male and female students, 40 male and female teachers, three educational managers, and seven parents. All the students in this study were registered in the last years of secondary middle school or just had taken part in the university entrance exam. Participants of the second phase of the study (the fifth group) were four elite in language teaching and assessment from Islamic Azad universities (three associate professors and one assistant professor) whose views were used in the impact analysis and structural interpretation of the model developed in the second phase of the research.

Procedure of the study

The primary phase of the study was carried out through some stages which are briefly discussed below:

Scale development

Development, validation, and administration of the scale were conducted in at least twelve steps which are discussed briefly below. These steps were conducted based on Dornyei (2010) on scale validation and Tabachnick and Fidell (2013) on structural equation modeling, including the following:

a) Interview with experts — aimed at identifying main items of the scale and steering the direction of the next step, i.e., literature review
b) Review of related literature — aimed at identifying the existing instruments, and models, establishing a theoretical framework for the research, and identifying components of the construct under study
c) Content selection — aimed at conducting content sampling and multi-item scales through content analysis of data collected in two previous steps;
d) d) Item accumulation and generation — aimed at covering most comprehensive content for the construct, one from the items of the scale will generate

e) Expert opinions and judgment — aimed at checking and controlling representativeness, accuracy, and intelligibility of the generated items
f) Designing rating scale — aimed at identifying the best design for scale administration. Here, the Likert scale was employed.
g) Designing demographic information for scale such as age, gender, education, and discipline
h) Designing the instruction for scale. This instruction ensures respondents about the anonymity of their responses. It also guides them on how to fill the questionnaire.
i) Initial piloting — a small-scale study aimed at investigating feasibility, time, cost, statistical variability, etc. The semi-structured interview was the primary tool used in scale development phase. Each interview session took 45 min to 1 h.

Scale validation

When the items were accumulated, the contents of the scales were selected, and primary questionnaire was formed, a pilot study was carried out, and possible linguistic mistakes, language ambiguities, etc., were detected and eliminated from the early scale. Then, the study underwent reliability estimation and validation of the model. For this purpose, a small-scale study was conducted to evaluate the feasibility of the research. Cronbach's alpha was selected among other measures of the reliability, to measure the internal consistency of the hypothesized model, and ensure that all items measure the same underlying construct in a consistent order. After obtaining an acceptable reliability index for the whole model and its subscales, the model underwent a validation process. Construct validation was carried out through a two-stage process in two separate administrations, including exploratory and confirmatory factor analyses. Exploratory factor analyses are used to identify main factors (components) and latent variables, while confirmatory factor analysis was used to verify the factor structure of the observed variables.

Interpretive Structural Analysis

Structural analysis, conducted in MICMAC software, is a semi-qualitative/semi-quantitative statistical procedure that allows the detection of a mutual influence and relationship between the factors of the studied system of test effects. According to Nazarko et al. (2017), structural analysis identifies the interaction between variables in a system to see whether and how factor X, for instance, exerts influence and or receives influence from factor Y, and answer whether factor X has a direct impact on the factor Y, and what is the intensity of this impact. In other words, the merit of the interpretive structural analysis is its ability to identify the ties between the variables, whose mutual influences are not obvious and may remain unrecognized even by experts in the field. Elaborating on this latest point, Godet (2008) holds that the structural analysis method helps reveal and describe mutual impact and the reaction based on which the researcher can determine which variables are crucial through analyzing dependencies between ostensibly irrelevant factors. It is also able to distinguish the variables that impact a given research area, including critical variables, target variables, determinant variables, regulatory, and supplementary variables, external variables, and strategic variables. These findings can feed subsequent interpretations of the developed model, including educational policy

makings aimed at modifying the testing system in the country and future studies aimed at predicting the future of language testing.

Results and discussion

As was indicated above, the first goal of the current study was the identification and validation of the construct of test consequences. Following Dornyei (2010) for scale development, and Tabachnick and Fidell (2013) for structural equation modeling, this objective was realized. To this aim, literature was reviewed, and a theoretical framework was established as a result of which content selection — aimed at content sampling and multi-item scales through content analysis proceeded, and five significant factors, as well as twenty-seven items, were identified. Then, to control the representativeness, accuracy, and intelligibility of the generated items, the accumulated items were investigated by the research team. In the next step, a Likert scale was designed. The early developed scale underwent a pilot study in which feasibility, time, cost, statistical variability, and internal consistency of the scale were calculated. The result of the internal consistency of the scale is presented in Table 1.

Also, items' total statistics were calculated, and then components' reliability indices were estimated. These results are reported in Table 2.

After estimating the reliability, the validation of findings was carried out to control constructs validity through appropriate measures, including factor loading and exploratory factor analysis. The newly developed scale was administered to the next group of participants, and exploratory factor analysis was carried out through SPSS.

In the earlier analyses, assumptions of exploratory factor analysis were investigated (Table 3) among which the assumptions of factorability of data, sample adequacy,

Table 1 Reliability statistics of the questionnaire

Cronbach's alpha	Cronbach's alpha based on standardized items	No. of items
0.89	0.911	27

Table 2 Components reliability indices

Factor	*R*
Teaching, learning, and assessment	0.90
Managerial	0.89
Attitude and perception	0.90
Emotion and self-confidence	0.89
Contextual	0.88

Table 3 Assumptions of factorability

Results of KMO and Bartlett's test		
KMO measure of sampling adequacy	0.796	
Bartlett's test of sphericity	App chi-square	3734.11
Df	121	
Sig	0.001	

and sphericity were more salient. According to Pallant (2007), the value of the Kaiser-Meyer-Olkin index ranges between 0 and 1, and the values above 0.6 are acceptable. Also, the value of Bartlett's test for sphericity needs to be below 0.05 (here, $p = 0.001$). Therefore, the data were appropriate for factorability (Pallant, 2007).

Also, the findings showed that under the umbrella term of test consequences, three significant factors and nineteen items were extracted which are shown in the following tables. Using Pallant (2013), items with unacceptable factor loading values were removed from the pool of items. The result of this rotation, based on Varimax rotation, is presented below (Table 4):

Thus, as it is shown in this table, three significant factors, as well as nineteen items (variables), were extracted which enjoy an acceptable loading factor value. The description of these factors and their items are presented in the following tables (Table 5).

In structural equation modeling, after identification of components and variables, model generation, aimed at depicting the relationship and interaction between variables and items in the scale, has proceeded. Accordingly, fit indices through confirmatory factor analysis were estimated. The findings of this stage are also presented below:

As it is revealed in Table 6, the chi-squared test (χ^2) is a statistic that represents the difference between observed and expected covariance matrices. Here, this value stands at 01.0, which shows an acceptable value and is thus suitable for a newly developed model. Also, the value of the normed fit index (NFI), which explores the discrepancy between the chi-squared value of the proposed model and the chi-squared value of the null model, is equal to 0.91, which shows a good fit. Besides, the relative fit index (RFI), which weights up the chi-square for the proposed model to a null model, is 0.93, which is acceptable. Furthermore, the value of the comparative fit index (CFI) that analyzes the model fit by examining the discrepancy between the data and the proposed model is

Table 4 Rotated matrix of factors based on PCA and Varimax

Item no.	1	2	3
Item 3	0.732		
Item 5	0.626		
Item 6	0.832		
Item 7	0.988		
Item 10	0.990		
Item 12	0.673		
Item 15	0.885		
Item 1		0.712	
Item 4		0.641	
Item 9		0.763	
Item 13		0.966	
Item 14		0.855	
Item 19		0.832	
Item 2			0.991
Item 8			0.969
Item 11			0.964
Item 16			0.994
Item 17			0.964
Item 18			0.978

Table 5 Descriptions of items extracted from exploratory factor analysis

Factor	Item	Description
Education and managerial	3	Assessment and feedback revolve around entrance exam
	5	Teaching methods are only consistent with entrance exam
	6	Content and material are only taught for the entrance exam
	7	Auditory and communicative skills are ignored for the entrance exam
	10	Teacher talk and interactions are centered on test-taking, grammar, and vocabulary
	12	Mother's tongue replaces foreign language
	15	The learning objectives of the entrance exam and language syllabus are inconsistent
Attitudinal and motivational	1	Testing competence importance prevails language competence for students
	4	Teaching for the test importance prevails teaching language skills for teachers
	9	Value of entrance exam performance prevails language learning for schools
	13	Students are more motivated to learn for the test rather than for language
	14	Teachers are more motivated to teach for the test rather than for the language
	19	Success is judged based on performance in entrance exams rather than language
Psychological and consequential	2	Iranian context both persuades for and forces test-taking rather than language
	8	Entrance exam exerts higher pressure, stress, and anxiety on students
	11	The entrance exam is more consistent with real-life needs
	16	The entrance exam is the only path for selecting the favorite educational channel
	17	The entrance exam is almost the only path to finding a job
	18	The entrance exam is the only path to acquiring social status

Table 6 Fit measures for model of test consequences

Index	Current level	Accepted level	Conclusion
Chi-square (χ^2)/df	1.0	< 3	Accepted
NFI	0.91	≥ 0.90	Accepted
RFI	0.93	≥ 0.90	Accepted
CFI	0.91	≥ 0.90	Accepted
RMSEA	0.01	< 0.05	Accepted

0.91, which is acceptable. Finally, the root-mean-square error of the approximation value was equal to 0.01, which is a fair index.

In the second phase of the research, structuring of findings and identifying the nature of the variables involved in the construct of test consequences were carried on through a matrix of crossed impact multiplications in MicMac software. Results of the analyses are shown below. In this phase of the study, the elite group of the participants rated the effect of items on each other in MicMac's characteristics matrix (Table 7).

Table 7 Matrix characteristics of MicMac components

Indicator	Value
Matrix size	19
Number of iterations	2
Number of zeros	51
Number of ones	29
Number of twos	80
Number of threes	176
Number of *P*	25
Total	310
Fillrate	85.87257%

In this table, matrix size shows the number of variables, 0s indicates no effect, 1s indicate weak effect, 2s indicate moderate effect, and 3s indicate strong effect while Ps indicate potential effect.

Figure 1 reveals a bulk of information on the structure and nature of variables in the construct of test consequences. The following information was extracted from the map:

A) *Stability/instability of the system*

According to the map of direct influence, the distribution of variables around axes in stable systems is L-shaped, while the scattered map of variables reveals that the system is unstable. In the current study, an L-shaped distribution of variables is not produced, meaning that the system of test consequences in Iran is unstable, and a change in one or two variables may bring about a drastic change in the map of the system. It might also change the position of each variable from one area to the other.

B) *Input variables (critical and environmental variables)*

Input variables are placed in the northwest area of the dependence/influence map (Fig. 1). These items are input variables of the system, i.e., change in them brings about change in other variables in the system. Input variables have two main kinds: (1) critical variables, which are the most important and influential variables in the system and manipulation of them exert critical effects on the system. According to the red diametrical line in the map, these variables are placed in the upper half of the northwest area of the map. In the present map, items 9, 19, 2, 16, and 17 are critical variables in the system of tests' consequences. (2) Environmental variables, these variables, despite their enormous effects on the system, are not open to change and manipulation. These variables are located in the lower half of the northwest area. In this study item, 18 falls in this category.

C) *Bidimensional variables (risk and target variables)*

These variables, which are placed in the northeast part of the map, are very unstable in a way that they can serve both as an input or output variable in different conditions. These variables fall into two significant categories:

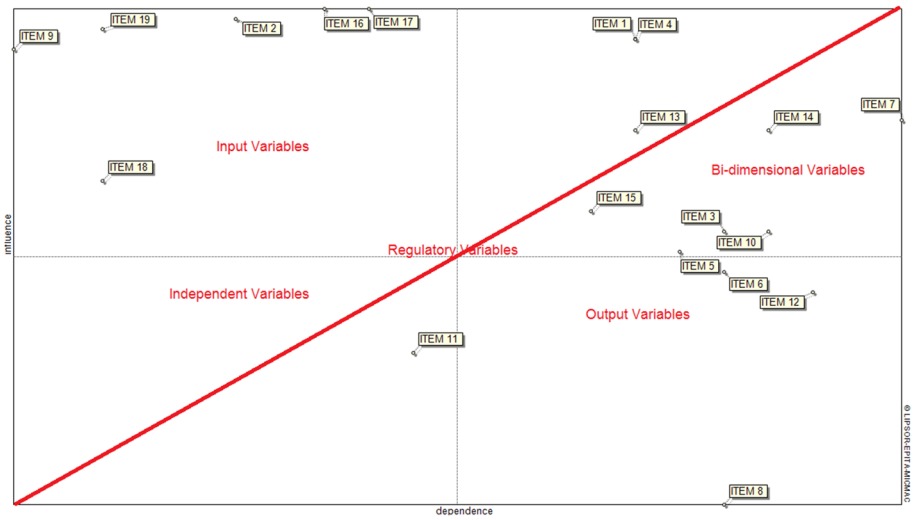

Fig. 1 Map of variables with direct influence/dependence relationship

1) Risk variables: due to their unstable nature, this category of variables can serve as the critical variables in the system and serve as influential input variables. With respect to the red diametrical line, these variables are located in the upper half of the northeast part of the map. In the current system, items 1, 4, and 13 fall in this category.
2) Target variables, unlike risk variables, these variables tend to serve more as an output variable — rather than an input variable — and receive effect and are located in the lower half of the northeast area of the map.

D) *Dependent variables (output variables)*

As their name suggests, these variables, which are placed in the southeast area of the map, are too sensitive to change in the variables explained above (B and C categories) and, therefore, serve as the output of the system. Items 6, 12, and 8 fall in this category.

E) *Independent (leverage) variables*

These variables which are placed in the southwest area of the map neither exert nor receive influence on/in the system since they are remotely attached to the system, and their relationship with the system is weaker than other variables in the current research. However, they tend to exert influence if manipulated. They can, therefore, be used as a secondary leverage variable to change the status of the system. In the current study, item 11 falls in this category.

In addition to information extracted from the map, the graphic schema of directions of strong effects between the variables of the study is represented in the following figure (Fig. 2). The lines of medium and weak effects are not drawn in the figure to avoid congestion in the relationship lines.

As it is revealed in Fig. 2, fourteen variables of the study have strong and direct effect on each other that submits further evidence on the sensitivity and instability of the construct of test consequences. The findings of this phase are in line with Frost (Frost: Test

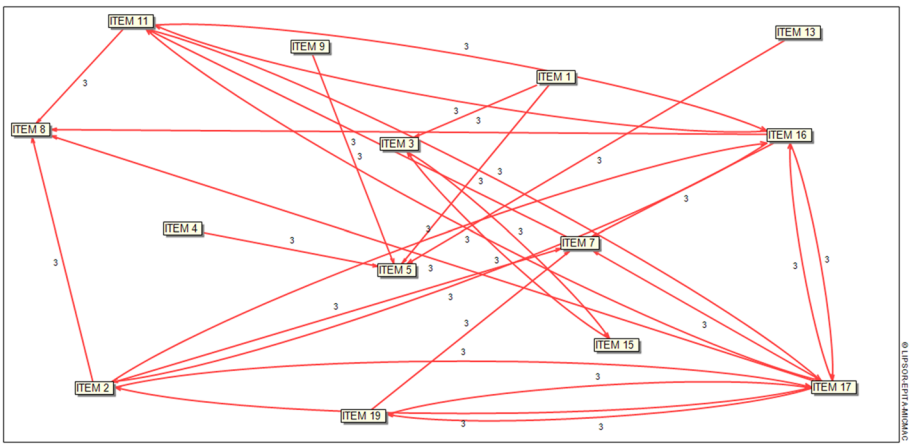

Fig. 2 Graphical representation of strong effects among variables of the study

impact as dynamic process: Individual experiences of the English test requirements for permanent skilled migration in Australia, unpublished), who had reported that test impact has a dynamic and complex nature because it embraces several abstract and concrete variables such as perception and attitude of stakeholders, the physical context of the study, content, and curriculum. Also, the findings of the current research are similar to those found in Hung and Huang (2019) which modeled nine main variables in the construct of washback. In addition to that, according to Li (2018), many alternative approaches to language testing are aimed at enforcing change in the curriculum, while without extracting the nature and weight of variables involved in test consequences, changing the curriculum is difficult. Like many methodologically robust studies which employed sophisticated statistics to develop a model of test influence, the present study went one step further and extracted the nature of variables involved in a validated construct of test consequences. This helps educational policymakers and practitioners to manipulate the input and critical variables to enforce change. In this respect, the current research is in harmony with Dimova (2017), depicting the inconsistency between educational policies and testing policies, stressing the divergence of intended and real uses of high-stakes tests. According to Cheng (1998) and Alderson and Wall (1993), the mismatch between curricular objectives and the real practice of teachers is most possibly related to the significant effect of the high-stakes tests such as university entrance exams on content of education and quality of teaching activities and practices. Also, the findings of the current study revealed a strong and overt impact of the university entrance exam on the education process in the Iranian context, which encompassed teaching, learning, and assessment. The findings in this section are in line with Ramezaney (2014), who verified the existence of a solid manifest washback impact of the English module of the university entrance exam on teachers' curricular planning and instruction. Furthermore, the findings of the current research support Pan and Roever (2016) which reported that high-stakes test consequences at both the micro-level of the individual test taker and the macro-level of society. Finally, these findings are in line with Cheng (2005) and McNamara (2000) who found that test impact exists in almost all those exams used for education, employment, promotion, immigration, citizenship, monitoring the

performance of schools and colleges, implementing educational policies, reforming educational systems, and deciding on the distribution of funding.

Having a look back at the literature reviewed, the findings of this study show close affinity with those of Di-Gennaro (2017) (Di-Gennaro: The Washback effects of an English exit exam on teachers and learners in a Korean university English program. Unpublished PhD dissertation), who had reported the effect of a high-stakes test in at least three dimensions, mainly on teaching and learning. Accordingly, when teachers in the study tended to use the textbooks and materials created by the department members, regardless of focusing on test-related items, paid little attention to the stakes of the exam. Furthermore, teachers tended to spend more time talking and lecturing in lower-level classes; however, direct preparation for exam items decreased as the level increased. Similarly, the language learners reported having changed the ways they studied English based on the high-stakes test rather than their regular assessments and reported that they are aware that their test scores are significant in their lives and that the test overall was very important and accordingly felt stressed or worried about the failure in the test. These findings stand robustly behind Ghorbani (2008), who reported on the weight and observed impact of the English module of IUEE on curricular planning of language education in the country since the research shed light on the strong impact of the exam on teaching and the perception of practitioners. The research had also discovered the impact of exams on several non-curricular issues related to the context of the study. Thus Ghorbani's results are in line with the findings of the current research that explored and confirmed the social and cultural dimension of washback in the model. Bachman and Palmer (2000) elaborate on the non-curricular effects of high-stakes tests and state that the intensity of the washback effect depends on the social and educational uses of the test to a large extent.

From a perceptional point of view, too, the findings of the current research stand behind Cheng et al. (2004) in that the perceived value and stakes of an exam can be very much dependent on social contexts and that the value placed upon the exam itself determines its stakes (Xie, 2015). Booth (2012) (Booth: Exploring the Washback of the TOEIC in South Korea. A sociocultural perspective on student test activity, Unpublished PhD Dissertation) has also found that test effects vary according to the status or level of the stakes of a test. This idea was previously arrived at by Alderson and Wall (1993) and Hughes (2003). Thus, it can be claimed that many of previous models and conceptualizations of test effects and consequences only investigated the influence of the community of test-takers themselves on the wide-scale impact and stakes of the tests while neglecting the social context and other stakes holders such as test consumers, teachers, and educational managers. In terms of perceptions and attitudes, the findings of the current research also endorse those of Ramezaney (2014) who had reported that participants of the study repeatedly demanded their teachers to use and explain items of the university entrance exam in their regular class practices because they had perceived that the university entrance is more important than their regular English class and acquisition of communicative competence. On the side of the teachers, too, the same inclination was reported by Ramezaney (2014), underscoring the overt zest of most teachers to prepare their students for the university entrance exam.

From an educational point of view, a validated model of test consequences in the country can provide educational practitioners and managers with a valid and credible source of information about the factors and the variables involved in the phenomenon. Also, the findings can help researchers to adopt a more targeted design and select the most important and appropriate variables to study. Findings from the structural interpretation of factors and variables can assist educational decision makers and policy developers to recognize and manipulate the most strategic variables interacting in the system of the test consequences. Any managerial scenario development and future study of the Iranian instruction and evaluation system also needs a thorough, practical, and extensive knowledge about the nature and behavior of the variables involved in the system. While most scenario and future studies use an at-hand complex of variables, the integration of SEM and structural interpretation provides the latter with statistically and scientifically validated input which is mostly related to the phenomenon being studied.

Conclusion

Considering procedural and methodological deficiencies in the literature on test consequences, the current study was an attempt to employ structural equation modeling and matrix of crossed impact multiplications and to shed light on the factors and variables of the English module of the IUEE, validate finding, and explore their nature and behavior in the system. The first research question asked about the main factors in IUEE's consequences. Exploratory factor analysis showed that three significant factors embracing nineteen variables are involved in test consequences. Also, the second question asked about the nature and behavior of these factors and variables. Analyses through MicMac showed that these variables are scattered around the matrix of influence/dependence to form an unstable system of input, output, independent, and bidimensional variables, each behaving differently in the construct of test consequences.

These findings have implications for research practitioners in humanities in general and language teaching and assessment in particular. Also, the results of the study are very informative to language teachers, learners, and, above all, language syllabus designers and educational policymakers. Primarily, the model extracted and interpreted above is helpful for language practitioners in Iran to identify alternative educational and assessment practices and promote educational system in the country. In addition to that, systematic analysis of test effects provides valuable information for the test developer and proposes a new methodological path for model development in washback and impact studies. Moreover, the findings of the structural interpretation and variable study is informative to educational managers, language policymakers, and scenario writers because the information obtained about the nature of the variables is a credible input for these processes. These information helps educational decision-makers to adopt the most effective strategies to eliminate or alleviate these consequences.

Though the research studied the complex phenomenon of test consequences from a new methodological perspective, some methodological, sampling, contextual, and even theoretical issues exist, which necessitate conducting further research. Many findings are drawn upon interdisciplinary areas such as psychology, sociology, administration, and management sciences, which share many with applied linguistics but still have severe limitations. Therefore, insights from further disciplines can be employed at the

level of data collection, data analysis, and discussion of findings. This study did not use a guided theoretical saturation in the data collection stage due to time limits and limitations in sampling procedures. Future research is required to collect data from different dimensions of the phenomenon until theoretical saturation in each stage. More rigorous tests and instruments can be used in data collection and data analysis. For instance, in the first stage of the study (scale development), to collect data from the literature, qualitative meta-analysis, documentary research, and systematic review can be employed. Besides, at the level of data analysis, for instance, more rigorous statistics and tests can be carried out. For example, in conducting interpretive structural modeling, in addition to the matrix of direct effects, results, interpretation, and graphs of the matrix of indirect effect can also be used. Since more rigorous statistics help generalize findings and validity of conclusions, future researchers are recommended to apply these suggestions in their succeeding research.

Abbreviations

CFA	Confirmatory factor analysis
CFI	Comparative fit index
EFL	English as foreign language
IUEE	Iranian universities' entrance exam
KMO	Kaiser-Meyer-Olkin
NFI	Normed fit index
RFI	Relative fit index
TEFL	Teaching English as a foreign language

Acknowledgements
I, on behalf of all authors, express deepest thanks to our colleagues in the Department of English (Faculty of Humanities, Islamic Azad University of Torbat Heydariyeh) who provided insight and expertise that greatly assisted our research.

Authors' contributions
All authors participated in designing, collecting information, analysis, and writing of this paper. The authors read and approved the final manuscript.

Funding
No funding received

Availability of data and materials
The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate
The current paper was approved by the Ethical Board of Iranian English language institutions. All participants provided written informed consent, and they fully understood the study purpose. Informed consent was obtained from all individual participants included in the study.

Competing interests
The authors declare that they have no competing interests.

Received: 13 June 2022 Accepted: 27 July 2022
Published online: 11 October 2022

References

Alavi, S. M., & Masjedlou, A. P. (2017). Construct under-representation and construct irrelevant variances on IELTS academic writing task 1: Is there any threat to validity? *Theory and Practice in Language Studies*, *7*(11), 1097. https://doi.org/10.17507/tpls.0711.19.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115–129. https://doi.org/10.1093/applin/14.2.115.

Ali, M. M., & Hamid, M. O. (2020). Teaching English to the test: Why does negative washback exist within secondary education in Bangladesh? *Language Assessment Quarterly*, *17*(2), 129–146. https://doi.org/10.1080/15434303.2020.1717495.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal, 2*, 1–34.
Bachman, L., & Palmer, A. (1996). *Language testing in practice*. Oxford University Press.
Bachman, L. F., & Palmer, A. S. (2000). *Language testing in practice*. Oxford: Oxford University Press.
Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice: Developing and using language assessments in the real world*, (2nd ed.,). Oxford University Press.
Bailey, J. L. (2018). A study of the washback effect of university entrance examination on teaching pedagogy and student learning behavior in Japanese high schools. *British Journal of Education, 6*(6), 50–72.
Bailey, K. M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, *13*(3), 257–279. https://doi.org/10.1177/026553229601300303.
Chapelle, C. A. (2020). An introduction to language testing's first virtual special issue: Investigating consequences of language test use. *Language Testing*, *37*(4), 638–645. https://doi.org/10.1177/0265532220928533.
Chen, Q., Hao, C., & Xiao, Y. (2020). When testing stakes are no longer high: Impact on the Chinese college English learners and their learning. *Language Testing in Asia*, *10*(1). https://doi.org/10.1186/s40468-020-00102-5.
Cheng, L. (1998). Impact of a public English examination change on students' perceptions and attitudes toward their English learning. *Studies in Educational Evaluation*, *24*(3), 279–301. https://doi.org/10.1016/s0191-491x(98)00018-2.
Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. Cambridge: Cambridge University Press.
Cheng, L. (2008). The key to success: English language testing in China. *Language Testing, 25*(1), 15–37.
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. https://doi.org/10.1017/s0261444815000233.
Cheng, L., Watanabe, Y., & Curtis, A. (2004). *Washback in language testing: research contexts and methods*. Routledge.
Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, *17*(1), 31–43. https://doi.org/10.1037/a0026975.
Davies, A. (1968). *Language testing symposium: A psycholinguistic approach*. Oxford University Press.
Dimova, S. (2017). Life after oral English certification: The consequences of the test of oral English proficiency for academic staff for EMI lecturers. *English for Specific Purposes*, *46*, 45–58. https://doi.org/10.1016/j.esp.2016.12.004.
Dong, M. (2020). Structural relationship between learners' perceptions of a test, learning practices, and learning outcomes: A study on the washback mechanism of a high-stakes test. *Studies in Educational Evaluation*, *64*, 100824. https://doi.org/10.1016/j.stueduc.2019.100824.
Dornyei, Z. (2010). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.
Estaji, M., & Ghiasvand, F. (2019). The washback effect of IELTS examination on EFL teachers' perceived sense of professional identity: Does IELTS related experience make a difference? *Journal of Modern Research in English Language Studies*, *6*(3), 103–183. https://doi.org/10.30479/jmrels.2019.11123.1391.
Farokhipour, S., Khoshsima, H., Sarani, A., & Ganji, M. (2020). Presenting and investigating the effect of a local model of dynamic assessment in diagnosing and removing learning difficulties of high school students in productive skills. *Foreign Language Research Journal*, *10*(1), 120–134. https://doi.org/10.22059/jflr.2019.273144.593.
Ghorbani, M. R. (2008). ELT in Iranian high schools in Iran, Malaysia, and Japan: Reflections on how tests influence the use of prescribed textbooks. *Reflections on English Language Teaching., 8*, 131–139.
Godet M. (2008). *La Prospective stratégique, pour les entreprises et les territories, Dunod*.
Green, A. (2007). *IELTS washback in context: Preparation for academic writing in higher education*. Cambridge University Press.
Green, A. (2020). Washback in language assessment. In C. A. Chapelle (Ed.), *Concise encyclopedia of applied linguistics*, (pp. 1159–1164). Wiley-Blackwell.
Hughes, A. (1989). *Testing for language teachers*. Cambridge University Press.
Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.
Hung, S. T. A., & Huang, H. T. D. (2019). Standardized proficiency tests in a campus-wide English curriculum: A washback study. *Language Testing in Asia*, *9*(21), 1–17. https://doi.org/10.1186/s40468-019-0096-5.
Jamalifar, G., Salehi, H., Tabatabaei, O., & Jafarigohar, M. (2021). Washback effect of the English proficiency test (EPT) on PhD candidates' language learning strategies. *Journal of Language and Translation*, *11*(3), 179–191. https://doi.org/10.30495/ttlt.2021.683485.
Kiani, G., Alibakhshi, G., & Akbari, R. (2009). On the consequential validity of ESP tests: A qualitative study in Iran. *Journal of English Language Pedagogy and Practice*, *2*(4), 103–126.
Kim, S. J. (2017). *Exploring a Framework for Consequential Validity for Performance-Based Assessments*. Retrieved from the University of Minnesota Digital Conservancy. https://hdl.handle.net/11299/191327.
Kuang, Q. (2020). A review of the washback of English language tests on classroom teaching. *English Language Teaching*, *13*(9), 10–17. https://doi.org/10.5539/elt.v13n9p10.
Kumaravadivelu, B. (2005). *Understanding language teaching: From method to post method*. Lawrence Earlbaum.
Larsson, M., & Olin-Scheller, C. (2020). Adaptation and resistance: Washback effects of the national test on upper secondary Swedish teaching. *The Curriculum Journal*, *31*(4), 687–703. https://doi.org/10.1002/curj.31.
Li, X. (2018). Self-assessment as 'assessment as learning' in translator and interpreter education: Validity and washback. *The Interpreter and Translator Trainer*, *12*(1), 48–67. https://doi.org/10.1080/1750399x.2017.1418581.
Mahmoudi, L. (2015). The washback effect of the Iranian national university entrance exam (INUEE) on pre-university students' English learning progress. *British journal of English Linguistics, 3*(3), 34–49.
McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.
McNamara, T. (2006). Validity in language testing: The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31–51. https://doi.org/10.1207/s15434311laq0301_3.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*, (3rd ed., pp. 13–103). American Council on Education/Macmillan.
Moss, P. A., Pullin, D., Gee, J. P., & Haertel, E. H. (2005). The idea of testing: Psychometric and sociocultural perspectives. *Measurement: Interdisciplinary Research & Perspective*, *3*(2), 63–83. https://doi.org/10.1207/s15366359mea0302_1.

Nazarko, J., Ejdys, J., Halicka, K., Nazarko, Ł, Kononiuk, A., & Olszewska, A. M. (2017). Structural analysis as an instrument for identifying critical drivers of technology development. *Procedia Engineering, 182*, 474–481.
Pallant, J. (2007). *SPSS survival manual: A step-by-step guide to data analysis using SPSS for Windows*. Open University Press & McGraw-Hill Education.
Pallant, J. (2013). *SPSS Survival Manual. A step by step guide to data analysis using SPSS* (4th ed.). Australia: Allen & Unwin.
Pan, Y. C., & Roever, C. (2016). Consequences of test use: A case study of employers' voice on the social impact of English certification exit requirements in Taiwan. *Language Testing in Asia*, *6*(1). https://doi.org/10.1186/s40468-016-0029-5.
Ramezaney, M. (2014). The washack effects of university entrance exam on Iranian EFL teachers' curricular planning and instruction techniques. *Procedia - Social and Behavioral Sciences*, *98*, 1508–1517. https://doi.org/10.1016/j.sbspro.2014.03.572.
Rezvani, R., & Sayyadi, A. (2016). Washback effects of the New Iranian TEFL Ph.D. Program entrance exam on EFL instructors' teaching methodology, class assessment, and syllabus design: A qualitative scrutiny. *Journal of Instruction and Evaluation*, *9*(33), 159–180.
Sadighi, S. (2018). Wash-back effect of Iranian students' pre-university English textbook and university entrance examinations: Teachers-based perspectives. *African Educational Research Journal*, *6*(4), 303–316. https://doi.org/10.30918/aerj.64.18.098.
Shih, C. M. (2007). A new washback model of students' learning. *The Canadian Modern Language Review*, *64*(1), 135–161. https://doi.org/10.3138/cmlr.64.1.135.
Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass, & C. Madden (Eds.), *Input and second language acquisition*, (pp. 235–253). Newbury House.
Tabachnick, B., & Fidell, L. (2013). *Using multivariate statistics*. Pearson Education Inc.
Tabatabaei, O., & Safikhani, A. (2011). The washback effect of university entrance exam on EFL teachers' methodology and test development. *Journal of English Studies*, *1*(2), 145–172.
Vernon, P. E. (1956). *The measurement of abilities*. University of London Press.
Xie, Q. (2015). Do component weighting and testing method affect time management and approaches to test preparation? A study on the washback mechanism. *System, 50*, 56–68.

Publisher's Note