Washback Effects of Board-Based Speaking Tests[1]

(Efectos colaterales de los exámenes orales por tribunal)

Henry Sevilla Morales[2]
Universidad Nacional, Costa Rica

Lindsay Chaves Fernández[3]
Universidad Nacional, Costa Rica

Abstract

This study analyzes the opinions of a group of three English as a Foreign Language programs, on the washback effects of board-based oral tests on the students' language proficiency and foreign language anxiety levels, and on the professors' instructional choices and decision making. With purposive sampling strategies and triangulation techniques, strong washback effects on learners' foreign language anxiety levels and professors' instructional choices, with lesser effects on learners' proficiency levels and instructors' decision making, were identified.

Resumen

El estudio analiza las opiniones de un grupo de estudiantes de tres programas de Inglés como Lengua Extranjera, sobre el efecto colateral de los exámenes orales por tribunal en su dominio lingüístico

1 Recibido: 14 de julio de 2019; aceptado: 11 de febrero de 2020. An earlier version of this study, "Washback of Board-Based Speaking Tests: Voices from the EFL Learner," was presented at the IV Conferencia Internacional de Lingüística Aplicada (CONLA-UNA), March 27-29, 2019, at the regional campus Brunca of the Universidad Nacional, Pérez Zeledón, Costa Rica.

2 Escuela de Literatura y Ciencias del Lenguaje. Correo electrónico: henry.sevilla.morales@una.ac.cr

3 Escuela de Literatura y Ciencias del Lenguaje. Correo electrónico: lindsay.chaves.fernandez@una.ac.cr

y niveles de ansiedad, y la metodología y la toma de decisiones por parte del profesor. Mediante técnicas de muestreo intencional y diversas técnicas de triangulación, se detecta un considerable efecto colateral de estos exámenes en los niveles de ansiedad de los estudiantes y en la metodología de clase, así como en su dominio lingüístico y la toma de decisiones por parte del profesor.

Keywords: English as a foreign language, language proficiency, instructional choices, oral exams
Palabras clave: Inglés como lengua extranjera, dominio lingüístico, metodología, exámenes orales

Introduction

According to professional literature on second language testing, *washback* (or *backwash*) can broadly be defined as "part of the impact a test may have on learners and teachers, on educational systems in general, and on society at large."[4] The interest in this subject area increased in the late 1980s, following the explosion of communicative language teaching (CLT) methodologies that had begun in the 1970s. From Hughes' 1989 publication of *Testing for Language Testers* to the present,[5] hundreds of studies have been undertaken to improve our understanding of the role tests play in the classroom, the curriculum, and society in the long run (see, for example, Alderson, 1990[6]; Alderson and Wall, 1992[7]; Bailey, 1999[8];

4 Arthur Hughes, "Achieving Beneficial Backwash," *Testing for Language Teachers* (Cambridge: Cambridge University Press, 2002): 53-57 (53). DOI: 10.1017/CBO9780511732980.007.

5 See Anthony Green, "Washback in Language Assessment," Achieving Beneficial Backwash," *International Journal of English Studies* 13, 2 (2013): 39-51. DOI: https://doi.org/10.6018/ijes.13.2.185891.

6 Charles J. Alderson, "Language Testing in the 1990s: How Far Have We Come? How Much Further Have We to Go?," Anivan Sarinee, ed., *Current Developments in Language Testing* (Singapore: SEAMEO, 1991): 1-26.

7 J. Charles Alderson and Dianne Wall, "Does Washback Exist?," *Applied Linguistics* 14, 2 (1993) 115–129. DOI: https://doi.org/10.1093/applin/14.2.115.

8 Kathleen M. Bailey, *Washback in Language Testing* (Princeton, NJ: Educational Testing Service, 1999).

Messik, 1996[9]; Cheng, Watanabe, and Curtis, 2004[10]; Shih, 2009[11]; Reynolds, 2010[12]; Gamboa and Sevilla, 2013, 2014[13]; and Mogapi, 2016[14]). Advances range from acknowledging the influence of tests in the classroom[15] to awareness of the need to align assessment and curriculum,[16] delimiting the scope of washback,[17] and identifying models for washback based on previous theoretical developments.[18]

Within this context, to date two main types of washback have been proposed: beneficial and negative. *Beneficial* washback deals with the positive effects of assessment in the educational landscape both in and outside the classroom; *negative* washback refers to detrimental effects resulting from poor evaluation practices which often lead to academic frustration, student and teacher anxiety, poor methodological choices, hasty curricular planning and decision making, and many others.

Despite these theoretical developments, reports from current literature suggest deficiencies which need to be addressed in order to devise research-based solutions to everyday pedagogical problems. Shih, for example, has proposed lack of empirical studies

9 Samuel Messik, "Validity and Washback in Language Testing," *Language Testing* 13, 3 (1996): 1-18. DOI: https://doi.org/10.1177/026553229601300302.

10 Lying Cheng, Yoshinori Watanabe and Andy Curtis. *Washback in Language Testing: Research Contexts and Methods* (Mahwah, NJ: Lawrence Erlbaum Associates, 2004).

11 Chih-Min Shih, "How Tests Change Teaching: A Model for Reference," *English Teaching: Practice and Critique* 8, 2 (2009): 188-206. <http://education.waikato.ac.nz/research/files/etpc/files/2009v8n2dial1.pdf>.

12 Jessica Reynolds, "An Exploratory Study of TOEFL students as Evaluators of Washback to the Learners," Research for the Master of Applied Linguistics, The University of Queensland (2010).

13 Roy Gamboa, and Henry Sevilla, "Assessment of Listening Comprehension in Public High Schools of Costa Rica: The West and Central Pacific Case," *Proceedings of the 13th International Conference on Education*, Honolulu, HA (2013): 1-45; Roy Gamboa and Henry Sevilla, "The Testing of Listening in Bilingual Secondary Schools of Costa Rica: Bridging Gaps Between Theory and Practice," *Revista Actualidades Investigativas en Educación* 14, 2 (2014): 1-23. DOI: https://doi.org/10.15517/aie.v14i2.14800.

14 Molefhe Mogapi, "Examinations Wash Back Effects: Challenges to the Criterion Referenced Assessment Model," *Journal of Education and e-Leaning Research* 3, 3 (2016): 78-86. DOI: 10.20448/journal.509/2016.3.3/509.3.78.86.

15 Bailey, 1.

16 Ana Muñoz and Marta E Álvarez, "Washback in an Oral Assessment System in the EFL Classroom," *Language Testing*, 27, 1 (2009): 1-17. DOI: 10.1177/0265532209347148.

17 Green, 40.

18 Shih, 188-206.

to corroborate commonly-held testing beliefs as a major drawback.[19] Along the same lines, other authors claim that much of the research on washback has focused on the influence of high-stakes assessment,[20] that is, evaluation "in which the results are likely to have a major impact on the lives of large numbers of individuals or on large programs,"[21] such as the TOEFL, IELTS or TOEIC tests. As a result, relatively less attention has been paid to research on the effects of low-stakes (i.e., classroom-based) testing on teaching and learning.[22] In the case of Costa Rica, studies of this kind can be reduced mainly to two unpublished inquiries: a final graduation project for a Masters' degree by Filomena[23] and a term paper by Fernández at the undergraduate level.[24] A few other researchers have addressed the topic but have not conducted systematic analyses of the actual washback effect of second language tests.

Where the current study was run—*Universidad Nacional* (UNA), Costa Rica—final *board-based tests* (BBTs)[25] were officialized in oral expression courses across various EFL programs in 2011. At that time, the goal was to help learners attain better English proficiency by having more than one professor evaluate their performance in different communicative tasks. Although faculty opinions in the English Department have been polarized as to the effects of this type of assessment in reaching that goal, no empirical studies

19 Shih, 189.
20 Muñoz and Álvarez, 1.
21 Christine Coombe, Keith Folse, and Nancy Hubley, *A Practical Guide to Assessing English Language Learners* (Ann Arbor: MI: The University of Michigan Press, 2007) xix.
22 Muñoz and Álvarez, 1.
23 Guisella Filomena, "The 2017 English Curriculum Reform in San José Night Academic High Schools: An Exploratory Study of Teachers' Perceptions," Final Graduation Project to obtain the Masters' Degree in Second Languages and Cultures, Universidad Nacional, Costa Rica (2017).
24 Jean Carlo Fernández, "Board-Based ('Tribunal') Oral Testing System Washback Effect: A Case Study at the University of Costa Rica, West Branch," Unpublished term paper for the course *IO-5600 Técnicas de Investigación*, Universidad de Costa Rica (2018): 1-26.
25 The term "board-based test" is the researchers' own translation for *exámenes por tribunal*, a scoring system commonly used in Costa Rican universities, where final oral exams in EFL speaking courses are prepared by the instructor but are administered and scored by an *examining board* composed of the course instructor plus two additional English instructors.

have yet been carried out to explore the effects of these tests on students, teachers, and instruction. Aware of the fact that a test alone can hardly boost language proficiency, in the present paper we attempt to fill this gap by offering preliminary evidence on the extent to which board-based oral tests can impact (1) students' language proficiency and foreign language anxiety levels, and (2) professors' instructional choices and decision making. The long-term purpose behind this inquiry is to survey the impact of this test and whether it is, in any measure, contributing to improved language competences in students. With this in mind, the research question guiding the study is: *In what ways are board-based final examinations in oral expression courses influencing students' language proficiency and foreign language anxiety levels, as well as instructors' methodological choices and decision making?*

For second language assessment, our research is relevant in three main ways: (1) It helps unveil preliminary perspectives on the issue, (2) it opens an avenue for future research on the topic, and (3) it yields baseline data for curricular decision making and planning, particularly for the accreditation processes that the university has been going through since 2006. Forthcoming studies should cover wider populations and contexts, as well as other methodologies as a basis for more solid insights into the topic. The following section provides an overview of previous studies informing this investigation.

Literature Review

Over the years, washback-related research has garnered the attention of many teaching disciplines including Science,[26] Medicine,[27]

26 Moses Orwe Onyango, "Assessment and Teaching Science," *The International Journal of Science* 15, 4 (2008): 255-259. DOI: https://doi.org/10.18848/1447-9494/CGP/v15i04/45709.

27 Mohammad Nases Shafiee Jafarabadi, Nazila Zarghi, Vahideh Zolfaghari, and Mohammad Reza Kargozari, "The Effect of Washback on Reading Comprehension of Medical Students in English for Specific Purposes Classes," *Future of Medical Education Journal* 4, 4 (2014): 28-31. DOI:

Mathematics,[28] and many others. When it comes to language instruction in particular, significant strides have been made from the 1990s to the present to define, problematize, and theorize on the effects that tests may have on language teaching and learning. To outline the evolution of this subject since its inception to the present, we offer a brief review of studies in the lines that follow.[29] To set the grounds for our review, a quick reference to the field of second language testing (SLT) in general is first issued. Then, a review of both empirical and research review articles is presented to show the chronological evolution of the language testing washback subfield.

As early as 1990, J. Charles Alderson claimed that whereas progress had been made in terms of testing content, methods, and analysis, little evidence existed that these developments translated into classroom practice[30]. His paper addresses a number of theoretical issues such as whether, and to what extent, progress has been made in SLT, whether test quality was better in 1990 than in the preceding years, how much understanding of the field had been achieved, and many other provoking questions aimed at problematizing the topic of SLT, of which washback was a central part in the article. At that time, Alderson concluded the following: (1) Recent research has started to challenge commonly-held views on language testing over the past two or three decades; (2) insights on test content and validity are questionable; and (3) "the apparent progress we think we have made – that we celebrate at conferences and seminars like this one, that we publish and publicise – may well not represent progress so much as activity, sometimes in decreasing circles"; however, the author is also confident that the foundations for effective testing have been established,

10.22038/FMEJ.2014.3604.

28 Luis J. Rodríguez-Muñiz, Patricia Díaz, Verónica Mier, and Pedro Alonso, "Washback Effect of University Entrance Exams in Applied Mathematics to Social Sciences," PLOS ONE 11, 12 (2016): 1-18. DOI: 10.1371/journal.pone.0167544.

29 The list of studies reviewed here is by no means exhaustive; readers can consult other studies for a broader view of the discipline's developments.

30 Alderson, 1-26.

and that future efforts can come into fruition if "patience" and "stamina" are combined in the SLT enterprise.[31]

Three years later, Alderson and Wall published a paper specifically on washback. They raised the question of whether washback really exists because, they argued, research on this area had traditionally focused on reported perceptions of how teachers believe that tests affect teaching, but not on actual observations of instruction and learning. They first look at the theoretical notions surrounding washback, survey empirical studies conducted on the subject, and finally propose new directions on the phenomenon. Suggested research areas include operationalizing the concept of washback (i.e., its scope, boundaries, and aspects of impact to include); considering findings from previous studies; moving from interviews and questionnaires to actual classroom observations, document analysis, and meetings with different educational actors; and triangulating the researchers' perceptions with those of the research subjects. Following this, in 1996 Messick published a paper titled "Validity and Washback in Language Testing."[32] Theoretical in nature, the article outlines noteworthy observations on the relation between test validity and washback. The author makes the trenchant point that evidence of the effects of testing in teaching and learning needs to be analyzed more carefully than we generally realize. Quite often, he argues, evidence of positive changes in teaching and learning take place after a test has been administered, but these can hardly be proven to result from the tests alone. In his own words, "washback is a consequence of testing that bears on validity only if it can be evidentially shown to be an effect of the test and not of other forces operative on the educational landscape."[33]

Toward the end of the 1990s, assessment expert Kathleen Bailey published "Washback in Language Testing," a critical review

31 Alderson, 24.
32 Messick, 1-18.
33 Messick, 2.

surveying the latest advancements in language testing washback.[34] The study begins by acknowledging a lack of research-based evidence to confirm the influence of tests upon instruction and learning, a point made previously by several authors such as Shohamy, Wall and Alderson, Andrews.[35] Then, it gives various definitions for washback and reviews research on the subject area. It contains three sections: "Research on Participants in the Washback Process," "Research on Processes and Products of Washback," and "Investigating Washback from the TOEFL 2000." The paper includes a comprehensive survey of research exploring washback and test-takers, washback and language teachers, and washback and other educational actors such as test designers, curricular authorities, materials developers and publishing companies, policy makers and many others. It also discusses methodological issues for investigating washback and concludes with suggestions on appropriate methodologies to use in future research on this area of inquiry.

At the turn of the century, Liying Cheng published a research review article aimed at sharing insights on the topic of washback from different perspectives, including general education and language teaching.[36] It also features a historical overview of the topic's development, current scope, and previous and present efforts to minimize the negative washback effects of tests. The study offers a solid account of previous studies and the status quo of washback in language assessment; however, one major limitation is that the author's conclusion is not based on the review presented throughout the paper. In a more theoretically solid paper, Vinson, Gibson, and Ross criticize the high-stakes testing policies of the United States from the pragmatic perspective of John Dewey and other education progressives.[37] The authors are blunt to label "testing regimes"

34 Bailey, 1-54.
35 Bailey, 1-2.
36 Cheng, Watanabe and Curtis, 1-34.
37 Kevin D. Vinson, Rich Gibson and E. Wayne Ross, "High-stakes Testing and Standardization: The Threat to Authenticity," Kathleen Kesson, ed., *Progressive Perspectives* (Burlington: University

as "a simplistic cure-all—an absolute panacea—" to the variously perceived issues and threats to the U.S. public education system.[38] Based on the work by John Dewey, their paper focuses on four main goals: (1) to argue that mandated standardized testing represented "little more than poor, absurdly disconnected, and uninspired pedagogy"; (2) to question common belief that true learning necessitates scores to prove itself effective; (3) to claim that these policies work against vigorous efforts to really improve schooling systems; and (4) to challenge the degree to which testing satisfies all the students' needs, especially those of speakers of English as a second language and those who come from marginalized social strata. These scholars conclude: "high-stakes standardized tests and test scores undermine high-quality education, genuine student/teacher motivation, and the benefits of diversity and inclusion."[39]

One year after the publication of Vinson et al.'s work, Lih-Mei Chen published an empirical study on the washback effects of a public exam on English teaching in Taiwan.[40] Guided by a relational method for research, the author surveyed and interviewed English teachers from a junior high school. Numerical data were analyzed through bivariate correlation and multiple regression analyses, while qualitative data were assessed for content analysis via a note-based technique. All in all, findings suggest that public examinations linked to educational reforms have an impact on curricular planning and instruction, but the impact on teachers is superficial since the content is influenced. However, the teaching methods remain unaltered due to lack of teacher training on how to align educational reforms with teaching methodologies.

Along the same chronological lines, Lying Cheng, Yoshinori Watanabe, and Andy Curtis published *Washback in Language*

of Vermont, 2001): 1-16. Available at: <http://richgibson.com/HighStakesTesting.htm>.

38 Vinson, Gibson, and Ross, 1.

39 Vinson, Gibson, and Ross, 1-2.

40 Lih-Mei Chen, "Washback of a Public Exam on English Teaching," The Ohio State University (2002).

Testing: Research Contexts and Methods,[41] a book discussing two core issues: concepts and methodologies, and research conducted in different parts of the world on the topic of washback. The authors cover an array of subtopics, such as the impact of testing on learning and instruction, methods used in washback studies, the link between washback and curriculum innovation, the impact of assessment-based reform on the teaching of writing in Washington State, washback of the IELTS test in New Zealand, Washback in class-room based assessment in Australia, teacher-related factors affecting washback, and many others. The book offers solid grounding for understanding the state of the art of this topic up to 2004.

Carolyn E. Turner then studied the perspectives of ESL secondary teachers when implementing educational innovations introduced via provincial examinations.[42] The study was conducted in Quebec (where French is the mother tongue and exposure to English is limited), as part of a larger, longitudinal investigation that looks into the effects of these reforms on teacher behavior and the instructional choice. Turner surveyed 153 ESL teachers to find out how they dealt with this high-stakes test. She found that most teachers report that they are willing to embrace the new testing system and to align curriculum, instruction, and assessment in general; however, evidence emerged that instructors grappled at times with conducting assessments for different purposes, such as classroom-based evaluations versus high-stakes provincial tests. The author concludes that teachers play a vital role in the educational reforms that any language program purports to implement.

One year later, Dina Tsagari reported on what has been done and what remains to be done on this subject area.[43] Outlining the main

41 Cheng, Watanabe and Curtis.

42 Carolyn E. Turner, "Professionalism and High-Stakes Tests: Teachers' Perspectives When Dealing with Educational Change Introduced Through Provincial Exams," *TESL Canada Journal* 23, 2 (2006): 54-76. DOI: https://doi.org/10.18806/tesl.v23i2.55.

43 Dina Tsagari, "Review of Washback in Language Testing: What Has Been Done? What More Needs Doing?", Lancaster University, Lancaster, UK (2007). Although this paper is an unpublished

theoretical issues and endeavors by prominent testing authors (e.g., J. C. Alderson, L. F. Bachman, K. M. Bailey, S. Messick, etc.), Tsagari discusses the washback of high-stakes examinations in language instruction and testing and in education as a whole. She concludes that although much progress has been made, a good deal remains to be done, which is perhaps best summarized by her quotation of Spratt, who points out the following:

> There is a need for more studies [on testing washback] to be carried out in different learning contexts. Use of parallel methodologies for studies in different contexts might also allow researchers to investigate some of the apparent contradictions in the findings to date.[44]

Two empirical studies from two distant geographical areas, Colombia and Taiwan, were published in 2009. In the Colombian paper, Muñoz and Álvarez analyzed the washback effect of a speaking assessment system on various aspects of language teaching and learning. Using a combination of qualitative and quantitative methods, the researchers conducted teacher and student surveys, class observations, and an outsider evaluation of student's oral performance. On the teaching side, the investigation looked at "congruence between curriculum objectives and instructional tasks," a "variety of assessment tasks and task design," and "detailed and specific feedback"; on the learning side, it explored the students' understanding of assessment criteria and their use of self-assessment as a way to develop criteria for success and foster autonomous learning.[45] In the Taiwanese study, Chih-Min Shih investigated the washback effects of the General English Proficiency Test (GEPT)

theoretical review posted on several sites by the author herself, it has been cited by a number of authors and includes a solid discussion of empirical research which is useful as theoretical background.

44 Tsagari, 58. Mary Spratt, "Washback and the Classroom: The Implications for Teaching and Learning of Studies of Washback from Exams," *Language Teaching Research* 9, 1 (2005): 5-29 (27).

45 Muñoz and Álvarez, 5-6.

on English teaching. Shih selected the applied foreign language departments of a university of technology (University A) and an institute of technology (University B). These universities were alike in a number of ways, their main difference being that University A did not require GEPT, whereas University B did. The researcher used comparative research designs, observations and interviews as data collection techniques. She also used participant selection criteria, as well as various triangulation strategies (e.g., triangulating the results from the observations of different classes with those gathered by interviewing teachers and students from these classes). The study's main finding was that "only courses which were linked to the departmental GEPT policy and whose objectives were to prepare students for the test were significantly affected."[46] Simply put, high degrees of washback were identified for courses which required direct preparation for the GEPT. After advising policy makers to consider teacher factors and micro-level contextual factors when using a test as a lever for change, Shih proposes a "new, tentative" model of washback to depict the effects of tests on teaching.[47]

In the following year, Jessica Reynolds wrote a postgraduate thesis on students' perspectives on the washback effects of the TOEFL test in three TOEFL preparation courses in the U.S.A. Data collection instruments included three semi-structured group interviews with each group participating in the study, a focus-group interview with the teachers of the preparation course, student surveys, and class observations. Roughly, the findings reveal a correlation between students' proficiency level and perceived negative washback; that is, "the more competent students were with English and the TOEFL, the more negative washback they perceived on their learning."[48] The author found that students were uncertain about the activities that best prepared them for the test, and about whether

46 Shih, 188.
47 Shih, 189.
48 Reynolds, iv.

preparing for the TOEFL and improving their English skills were conflicting or mutually complementary undertakings.

That same year Annela Teemant explored students' "opinions, concerns, strategies, and preferences in testing" in general.[49] Subjects included six female and seven male Belorussian, Russian, Portuguese, Spanish, Korean, Chinese, and Arabic ESL students who had lived in the U.S.A. between 7 months and 14 years and had been college students between 1 month and 2.5 years. The purpose of this qualitative analysis was to become familiar with students' viewpoints on classroom assessment practices and to identify how faculty can address students' testing concerns. The author concludes that more research is needed to "clarify the degree of importance such factors as language proficiency, test foreign language anxiety, and format preferences play in ESL students' performance."[50]

In 2012, Melor Md Yunus and Hadi Salehi looked at the washback effect of an admission test—called the *Entrance Exam of the Universities (EEU*)—for higher education in Iran.[51] To this end, thirty pre-university students and 36 high-school instructors were randomly selected to assess the impact of this test on students' learning of English. Findings revealed that the exam had a negative influence on English instruction, directing it toward the grammar-based contents of the EEU and therefore neglecting speaking, writing, and listening from classroom instruction. The authors recommend a revision of the EEU's format and further research on the perceptions of stakeholders on this high-stakes test.

In 2013, a theoretical review article by Anthony Green surveyed the progress made around this topic since Hughes' *Testing for Language Testers* (see introduction) in 1989. Here, the author offers an

49 Annela Teemant, "ESL Student Perspectives on University Classroom Testing Practices," *Journal of the Scholarship of Teaching and Learning* 10, 3 (2010): 89-105.

50 Teemant, 101.

51 Melor Md Yunus, and Hadi Salehi, "The Washback Effect of the Entrance Exam of the Universities (EEU) on the Iranian Pre-university Students' English Learning," *The International Journal of Learning* 18, 7 (2012): 101-125. DOI: https://doi.org/10.18848/1447-9494/CGP/v18i07/4766.

extended definition for washback, discusses several relevant aspects of it, reviews research on it, shows how research has prompted the development of theoretical models of washback,[52] and suggests ways for test designers to take into account the washback effects of the tests they create. One major conclusion is that greater involvement of administrative authorities, textbook producers, instructors, and even students in the test development process could help bring teaching and instruction together.[53] Along the same lines, Green asserts that current testing practices can be enhanced through research evidence, and that washback needs to be studied and understood within "specific contexts of test use." The author is optimistic that a good deal of progress has been made around this subject, but he is also critical that little is known about the roles of various educational actors in the generation of test washback, with students being, perhaps, "the most important participants of all."[54]

Two more studies were published that year on testing washback: one from Taiwan and one from Cyprus. In Taiwan, Yi-Ching Pan studied whether and to what extent English certification exit requirements such as the GEPT, TOEFL, TOEIC, and IELTS tests have motivated a teaching-to-the-test modus operandi in tertiary colleges and universities, or if, on the contrary, these have geared teachers toward integrating the four macro skills of the language to attain communicative competence. Data were collected from teachers' questionnaires, teacher interviews, and classroom observations across various high schools in Taiwan. Findings indicate that these exit requirements have exerted little influence on classroom dynamics. They also suggest that these requirements did not encourage the teaching of communicative-oriented competencies.[55] The author warns, however, that if these exams continue to be compulsory

52 See, for example, Shih's 2009 paper reviewed above.
53 Green, 49.
54 Green, 49.
55 Yi-Ching Pan, "Does Teaching to the Test Exist? A Case Study of Teacher Washback in Taiwan," *The Journal of ASIA TEFL* 10, 4 (2013): 185-213.

in Taiwan, efforts must be made to develop test tactics without reducing instruction to the passing of a high-stakes examination.[56] In Cyprus, Georgia Vraketta explored the impact of the Pancyprian Examinations on instructors' stress levels. Using a mixed-methods design and snowball sampling procedures, the researcher examined the views of teachers from the private, public, and higher education settings on the negative washback brought about by the test.[57] Roughly, findings identified time allocated for grading, lack of time to cover the test syllabus, the students' parents, and the educational system of Cyprus as the main stress-generating factors.

The following year, Yi-Ching Pan published a study measuring Alderson and Wall's hypothesis that "a test would influence 1) degree/depth of learning, 2) attitudes toward methods of learning, and 3) some learners but not others."[58] Based on the premise that compared to teacher-related washback studies, research on student washback remains limited, Pan studied a cohort of 589 students from a technical university which held an exit test policy in Taiwan. The author compared student views before and after taking two exit exams, namely the TOEIC and the GEPT. Findings suggest that exit requirements yield different washback on different students depending on their "years of study, proficiency levels, and perceptions of tests."[59]

A summary of what had been achieved up to 2015 about washback was published in "Language Testing: The State of the Art," an article deriving from an interview with renowned assessment scholar James Dean Brown.[60] Here, among other aspects Brown gives a definition for language testing, reflects upon differences between assessment

56 Pan (2013) 201-202.
57 Vraketta, 26.
58 Yi-Ching Pan, "Learner Washback Variability in Standardized Exit Tests," *The Electronic Journal for English as a Second Language* 18, 2 (2014): 1-30.
59 Pan (2014) 21-22.
60 James Dean Brown and Salamani Nodoushan, "Language Testing: The State of the Art," *International Journal of Language Studies* 9, 4 (2015): 133-143.

and evaluation, lays out his views on high-stakes and low-stakes testing, and critiques the implications of language testing for social policy. He is especially critical of over-relying on the native speaker (NS) model to assess language proficiency, which perpetuates attitudes of inferiority toward anything that does not approximate to the NS standard.[61] On the whole, he views language assessment as a process where positivist perspectives conflict with postmodernist interpretivist views, as well as with "everything [else] in between."[62] He concludes the interview by giving experiential advice to younger researchers from the testing subfield, such as following their research interests, working systematically, asking constant questions, and enjoying the work they do.

In 2016, Molefhe Mogapi studied the washback of an examination system for primary school levels in Botswana, a country that transitioned from norm-referenced to criterion-referenced testing in 1994. Using qualitative and quantitative methods, the author surveyed the opinions of 66 "practicing" teacher students from the Department of Primary Education at the University of Botswana across ten teaching districts of the country.[63] Findings suggest a possible negative washback of these examinations on teaching and learning, with the narrowing of the syllabus being the most salient negative effect, thus indicating that tests fail to include vital elements from the syllabus.[64] The author also highlights the teaching-to-the-test effects introduced largely by testing booklets, which is working against Botswana's criterion-referenced intended policies.

Two papers from the Middle East offer a glimpse of the latest empirical research on washback. In Saudi Arabia, Abduljalil Nasr Hazaea and Yayha Ameen Tayeb investigated the washback of the Learning Outcome Based English Language Assessment approach

61 Brown, 137.
62 Brown, 133-134.
63 Mogapi, 78-86.
64 Mogapi, 85-86.

(LOBELA) on teaching methods, content assessment, teachers' attitudes and motivation.[65] Using a mixed-methods approach to research, the authors surveyed 36 lecturers from Najran University and interviewed 13 of them about their views on the LOBELA's washback effects. According to the subjects, the greater negative effect was on teaching methods, followed by content assessment, instructors' attitudes, and their motivation, respectively.[66] In Iran, Kioumars Razavipour, Sayyed Rahim Moosavinia, and Somayyeh Atayi studied the washback effect of the English Literature Module of the Admission Test of English Literature (ATEL) toward learners' attitudes and test preparation resources.[67] The sample included 100 graduate students taking their M.A. in English literature at eight Iranian state universities. Data were analyzed with descriptive statistics, thus revealing that the test affected participants' attitudes and their learning of English literature.[68]

From the studies reviewed above, a number of relevant observations can be drawn. Firstly, a good deal of the scholarly discussion on washback has been based on theoretical reviews and monographs from renowned authors such as Kathleen Bailey, James Dean Brown, J. Charles Alderson, Diane Wall, and many others, with a lack of empirical studies to test the validity of these authors' arguments. Secondly, the birth of the 21st century saw the emergence of two types of washback studies: those focusing on students' perspectives and those dealing with teachers' standpoints, with a reported prevalence of teacher-related washback over student-based

65 Abduljalil Nasr Hazaea and Yayha Ameen Tayeb, "Washback Effect of LOBELA on EFL Teaching at Preparatory Year of Najran University," *International Journal of Humanities and Applied Social Science (IJHASS)* 3, 3 (2018): 1-14.

66 Hazaea and Tayeb, 12.

67 Kioumars Razavipour, Sayyed Rahim Moosavinia and Somayyeh Atayi, "Construct Ambiguity and Test Difficulty Generate Negative Washback: The Case of Admission Test of English Literature to Graduate Programs in Iran," *International Journal of Instruction* 11, 4 (2018): 717-732. DOI: https://doi.org/10.12973/iji.2018.11445a.

68 Razavipour, Moosavinia, and Atayi, 717.

washback studies (see, for example, Pan).[69] Thirdly, as Muñoz and Álvarez have noted, the research agenda of recent decades has been dominated by the washback of high-stakes tests, which calls for more investigation on the effects of "classroom-based assessment on instructional and learning practices."[70] Lastly, the majority of washback-related research in ELT has centered on ESL, rather than on EFL contexts.

Arguably, the current research helps fill some of these gaps by: 1) providing empirical evidence for future EFL teachers who will soon be preparing and administering assessments; 2) providing student-based data on their views about the washback of board-based oral tests at a classroom level; and 3) reaching a more balanced state of the art by adding to the body of empirical EFL studies on the subject.

Methodology

Research Approach and Sampling Procedures

The current inquiry is subject to various classifications according to its depth, purpose, scope, and design. In terms of depth, it is exploratory since it seeks to unveil preliminary findings on the phenomenon studied and prepares the grounds for future studies;[71] regarding purpose, it is conceived as basic research because for now its purpose is to develop theory which can be applied in future decision making and curricular planning;[72] with regard to scope, it is cross-sectional since it studied a phenomenon over a short period of time (one year).[73] In terms of design, the investigation adopts the

69 Pan (2014) 1.
70 Muñoz and Álvarez, 1.
71 Roberto Hernández Sampieri, Carlos Fernández Collado and María del Pilar Baptista Lucio, *Metodología de la investigación,* 5th ed. (Mexico D. F.: McGraw-Hill, 2010) 77.
72 Larry R. Gay, Geoffrey E. Mills and Peter Airasian, *Educational Research: Competencies for Analysis and Applications*, 9th Ed. (Upper Saddle River, NJ: Pearson, 2009) 17.
73 Kate Ann Levin, "Study Design III: Cross-sectional Studies," *Evidence-Based Dentistry* 7, 1 (2006): 24-25. DOI: 10.1038/sj.ebd.6400375.

convergent parallel mixed methods design, a model where, in Creswell's words, "a researcher collects both quantitative and qualitative data, analyzes them separately, and then compares the results to see if the results confirm or disconfirm each other."[74]

The analysis was based on a side-by-side comparison of the datasets gathered, with a report on the numerical results first, and a qualitative account confirming or disconfirming such statistical results afterwards.[75] Following the analysis, a global interpretation of the impact of BBTs was issued in the form of a discussion section, and findings were then theorized in the light of the studies presented in the *literature review* section.

Participants and Context

Participants included 100 EFL-UNA students randomly selected from the three B.A. programs of the *Escuela de Literatura y Ciencias del Lenguaje* (ELCL, School of Literature and Language Sciences): B.A. in English, B.A. in English Teaching for Secondary Schools, and B.A. in English Teaching for Primary Schools. Each program covers over 40 courses (around 140 credits) spread throughout 8 academic semesters lasting 17 weeks. For the two B.A. programs focusing on teaching, courses touch upon the following axes: literature, linguistics, intercultural communication, writing, oral expression and listening comprehension, pronunciation, grammar, pedagogy and TESOL, humanities, and a number of electives. In the B.A. in English, the pedagogy and TESOL component is left out and replaced with translation and interpretation-based courses. With proficiency levels ranging from A2 in the first semester (freshman year) to C1 toward the end of the program (senior year), informants take between 6 and 9 oral expression courses framed within a

74 John Creswell, *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 4th ed. (Thousand Oaks, CA: SAGE, 2014) 219.

75 Cresswell, 222.

series of thematic areas such as society and humanism, science and technology, elocution skills, and commerce and economy. For purposes of scope, the current paper deals exclusively with the students' views on the elements studied (see below).

The data collection process took place in two phases over the course of 2018. The first phase was carried out during the first academic semester, with 67 freshmen from these programs being surveyed for their perceptions of the board-based oral exam on two areas: (1) students' language proficiency and foreign language anxiety levels, and (2) their professors' instructional choices and decision making.[76] On conducting a preliminary analysis of the results, the researchers decided to expand the scope of the study by surveying 33 more students from the three remaining levels (sophomore, junior, and senior) to contrast their views with those of freshmen. Both groups of informants were contacted in person as soon as they had finished their final board-based exam at the end of the first and second semesters of 2018, respectively.

The Board-Based Test

As stated above, the BBT became an assessment requirement for all oral expression courses in 2011. As such, learners are aware of the test since the first day of class when the instructor goes over the evaluation percentages in the syllabus. The topics covered in this test are studied throughout the semester in different individual and group activities in which the students learn vocabulary and expressions that enable them to speak accurately about the topics selected for the course. At the end of the semester, during evaluation week, the learners take the test in pairs or small groups, generally by randomly choosing one or two teacher-created questions related to the

76 It should be noted that we do not necessarily expect such a connection to exist (cf. introduction of this paper); our inquiry simply looks at whether the 2011 agreement is actually rendering the results it purported to render.

topics studied in class. Once the students select the questions, they speak about them for approximately 10 minutes, and while they do so, their conversation is recorded to leave an audit trail, and thus safeguard assessment decision-making in case students question their score. During test administration, the examiners listen and take notes of students' performance and, if necessary, ask follow-up questions. Prior to the test, testees are given a teacher-developed rubric indicating the intended performance criteria, which usually include grammar, pronunciation, content, and vocabulary. On the day of the test, three professors evaluate the learners' performance separately and allocate their grades either consensually or individually. BBT percentages generally range from 25 to 30 of the final grade in the course.

Instrument

The instrument consisted of a bipolar Likert scale of agreement based on Menold and Bogner's discussion on scale polarity design.[77] It measured the students' perceptions on the two areas examined (see above) through a series of statements to be assessed via opposite continua that dealt with the following response categories: level of agreement (*strongly agree/strongly disagree*), level of quality (*excellent/poor*), level of concern (*extremely concerned/not at all concerned*), level of problem (a serious problem/not a problem at all), knowledge of action (*never true/always true*), and likelihood (*extremely likely/extremely unlikely*). The scale contained the research objective, general instructions, and a list of statements for informants to assess in terms of the response categories (see appendix 1). In line with the recommendations by Menold and Bogner, it used five-point response categories in most statements, except for seven and

77 Natalja Menold and Kathrin Bogner, "Design of Rating Scales in Questionnaires," *GESIS Survey Guidelines* (2016): 1-13. DOI: 10.15465/gesis-sg_en_015.

four points in two statements with special characteristics.[78] Lastly, to capture the informants' qualitative assessments on the phenomenon, a space was provided for comments which were used in the side-by-side comparison to cross-check the numerical findings.

Ethics and Validity Measures

To conform to research ethics, a detailed letter of consent was signed by each participant. They also received detailed explanations on their rights as informants and the voluntary nature of their participation. Citation codes were assigned in the analysis of the qualitative datasets to protect their identities. A series of measures were taken to guarantee internal and external validity. Internal validity was achieved by pilot-testing the instrument with 23 students and six professors from the program prior to administration, and by employing random sampling to avoid selection bias.[79] For external validity, the researchers recruited participants who were not enrolled in their own classes at the time of the study, to prevent the *Hawthorne effect*, understood here as participants' altering of responses due to the psychological effects produced by the mere participation in research.[80] In addition to this, triangulation was used at two levels: researcher triangulation (two researchers contributing to findings) and theoretical triangulation (discussion articulated around previous studies reviewed). An audit trail was also left in case any member of the research community wishes to verify the accuracy of the results reported.

As a basis for the data analysis, attention should be given to operational definitions for the variables in the two areas investigated. For the purposes of this study, *English proficiency level* is understood simply as students' ability to use spoken English skills to communicate fluently and accurately according to the communicative event given.

78 Menold and Bogner, 2.

79 Louis Cohen, Lawrence Manion, and Keith Morrison, *Research Methods in Education*, 7th ed. (New York: Routledge, 2011) 184.

80 Cohen, Manion, and Morrison, 186.

Foreign language anxiety is defined as the feelings of concern, stress, and tension that arise while communicating in a foreign language, especially in classroom settings.[81] Instructional (or methodological) choices were conceptualized as the range of strategies, techniques, and activities used to teach a particular topic.[82] Decision making was operationalized as the degree to which the professor uses test results to adjust his/her planning, teaching, and assessment methods, as Richards and Lockhart have suggested.[83]

Data Analysis

This section contrasts the quantitative and qualitative data collected from the two groups of informants: (a) freshmen versus (b) sophomores, juniors and seniors. The data will be displayed by variables investigated in the form of a comparative analysis between the two groups. Numerical data will be presented concurrently with qualitative comments to provide further detail on the percentages obtained. Due to space constrictions, only the most prevalent data will be analyzed, and citation codes (from P-1 to P-100) will be used to refer to participants' comments.

The Test and Students' Language Proficiency

For freshmen, the BBT has exerted a moderate influence on their proficiency level since for the question of whether the test had helped them improve their general speaking skills in English, 28.4%

81 Şenel Elaldı, "Foreign Language Anxiety of Students Studying English Language and Literature: A Sample from Turkey," *Educational Research and Reviews* 11, 16 (2016): 219-228. DOI: 10.5897/ERR2015.2507.

82 Ian Tudor, *The Dynamics of the Language Classroom* (Cambridge: Cambridge University Press, 2001) 34.

83 C. Richards, and C. Lockhart, *Reflective Teaching in Second language Classrooms* (Cambridge: Cambridge University Press, 1994) 78. Qtd. in Eri Osada, "A Teacher's Decision-Making Process in an Elementary School EFL Education," *International Journal for 21st Century Education* 3, 2 (2016), 17. DOI: https://doi.org/10.21071/ij21ce.v3i2.5851.

reported to agree strongly and 34.3% said they simply agreed, as shown in table 1.

Table 1. Influence of Test on Freshmen's Speaking Skills

Descriptors and Percentages					Total
Strongly Agree	Agree	Neither or N/A	Disagree	Strongly Disagree	67
28.4%	34.3%	29.9%	3.0%	4.5%	100%

Table 2 displays the percentages regarding the quality of speaking skills prompted by the course contents according to students. As can be observed, the majority rated it as excellent, followed by a smaller percentage that perceived it as very good. None of learners assessed it as poor.

Table 2. Influence of Course Contents on Freshmen's Speaking Skills

Descriptors and Percentages					Total
Excellent	Very Good	Good	Fair	Poor	67
47.8%	38.8%	11.9%	1.5%	0%	100%

In addition, the qualitative annotations reveal concerns of various kinds. Participant 1, for example, suggested that "it would be great to have several "Artificial Tribunal" before the main one" (sic), similar to P-2, who wrote: "Maybe if the professors make a practice of the tribunal exam, that help a little bit more" (sic), and to P-30, who is concerned about his/her classmates feelings: "Personally, I don't mind tribunal tests, but I think it makes my classmates really nervous." Others, such as P-24, bluntly criticized that "the board-based test doesn't help to that." Yet, others argued that the test was necessary as training for the challenges of their future professional careers: "It make me feel nervous but that's part of what I am going

to have in a classroom so in my personal opinion I think it is good" (P-47, sic); "it has been very useful because it was a good practice to avoid the fear factor when talking in front of people" (P-50, sic).

Table 3 shows how for sophomores, juniors, and seniors, the test's contribution to their proficiency level is less evident. As an illustration, 45.5% disagreed that the exam influenced their general speaking skills, while only 12.1% agreed that it did so.

Table 3. Influence of Test on Sophomores, Juniors, and Seniors' Speaking Skills

Descriptors and Percentages					Total
Strongly Agree	Agree	Neither or N/A	Disagree	Strongly Disagree	33
0%	12.1%	33.3%	45.5%	9.1%	100%

Table 4 deals with the quality of speaking skills fostered by the contents studied; 48.5% agree that the contents studied in class indeed helped them improve their speaking skills. However, 36.4% expressed the contents did not.

Table 4. Influence of Course Contents on Sophomores, Juniors, and Seniors' Speaking Skills

Descriptors and Percentages					Total
Strongly Agree	Agree	Neither or N/A	Disagree	Strongly Disagree	33
9.1%	48.5%	36.4%	6.1%	0%	100%

Qualitatively, the comments are even more thought-provoking. For instance, P-77 questions the threatening nature of the exam: "This type of tests is super intimidating and causes a lot of foreign language anxiety to students and being nervous can affect your grade" (P-77, sic); P-80 questions its validity: "Although students share their current language development, a short interaction with

a partner in a "tribunal" test will provide little or no progress"; just as P-84 does when s/he highlights that "students tend to be really nervous and tend to forget the words or just pronounce incorrectly. It [the tests] is not measuring the real knowledge" (sic). To better capture the impressions on this variable, the following comments are presented below:

> "Most of the time I get nervous when I'm in a board-based, so I don't do my best." (P-97)
> "[…] having people judging me decrease the chances of showing my five speaking skills." (P-99, sic)
> "I think that the process helps but the exam doesn't." (P-92)
> "I believe that this test only evaluates how students develop in that day at that time and not the process we have had." (P-86, sic)
> "The pressure in those tests sometimes limits the performance of the students. Usually, presentations work best." (P-100)
> "Additional oral expression resources must be provided in order ro further develop speaking proficiency." (P-80, sic)
> "[The contents are] too narrow." (P-81)

Overall, the groups sampled responded differently on how they interpret the BBT's influence on their overall language proficiency. How the students assess its impact on their foreign language anxiety levels is analyzed below.

The Test and Students' Foreign Language Anxiety Levels

When dealing with foreign language anxiety, freshmen appear to perceive different levels of anxiety both *before* and *after* taking the evaluation (see tables 5 and 6). When asked about their feelings prior to the test, most of them agreed on having anxiety levels, as a total of 61.2% claimed to feel extremely and moderately concerned. The full range of responses is shown in table 5.

Table 5. Freshmen's Level of Concern *before* the Test

Descriptors and Percentages					Total
Extremely concerned	Moderately concerned	Somewhat concerned	Slightly concerned	Not at all concerned	67
29.9%	31.3%	20.9%	11.9%	6.0%	100%

Some of the reasons reported about why they felt anxious before the BBT had to do with their lack of control over the situation itself. That can be exemplified by the following comments: "because I couldn't choose my partner" (P-10), "it was the first time doing that kind of test" (P-13), and "I was intimidated by the presence of three professors" (P-17).

After taking the test, however, students' foreign language anxiety levels dropped considerably since, as seen in table 6, 20.9% affirmed not to feel concerned at all and only 9% claimed to feel extremely concerned. However, foreign language anxiety levels were still an issue for 28.4% of the students who claimed to feel moderately concerned after taking the test.

Table 6. Freshmen's Level of Concern *after* the Test

Descriptors and Percentages					Total
Extremely concerned	Moderately concerned	Somewhat concerned	Slightly concerned	Not at all concerned	67
9%	28.4%	20.9%	20.9%	20.9%	100%

Students' comments, on the other hand, show that they were, by and large, worried about having done well on the test, but they were aware of the fact they had not achieved a perfect mark. For instance, P-41stated: "I think that I was good and I hope so" (sic), whereas P-46 reported: "I know that I did right but we all do mistakes" (sic).

Table 7 displays freshmen's views on what BBTs represent for

them. Most seem to be almost certain that BBTs represent a moderate (32.8%) or minor problem (29.9%). They claim to feel concerned about factors such as final scores, their classmates' competence level, progress achieved, readiness for the test, among others.

Table 7. Freshmen's Degree of Concern about the Test

Descriptors and Percentages					Total
A serious problem	A moderate problem	A minor problem	Not a problem at all	No Answer	67
7.5%	32.8%	29.9%	22.4%	7.5%	100%

Like the freshmen, a majority of sophomores, juniors, and seniors (66.7%) expressed that their foreign language anxiety levels were quite high before the administration of the test. Table 8 shows that 21.2% informants described themselves as being extremely concerned and 45.5% as being moderately concerned. On the qualitative side, P-88 stated: "It usually makes me feel scared or something," and P-77 expressed "Imagine that you have to give a test in front of three important professors, I was obviously concerned."

Table 8. Sophomores, Juniors, and Seniors' Level of Concern *before* the Test

Descriptors and Percentages					Total
Extremely concerned	Moderately concerned	Somewhat concerned	Slightly concerned	Not at all concerned	33
21.2%	45.5%	15.2%	15.2%	3%	100%

However, the figures change for these two groups of participants when dealing with their post-test feelings (see tables 6 and 9), since 6.1% of sophomores, juniors, and seniors expressed feeling extremely concerned, as opposed to only 20.9% of freshmen who claimed not to be concerned at all. Most sophomores, juniors and seniors affirmed to feel moderately (21.2%) or somewhat (36.4%)

concerned after the BBT due to factors such as grades, the type of questions they would be asked, and the lack of professors' interaction during test administration.

Table 9. Sophomores, Juniors, and Seniors' Level of Concern *after* the Test

Descriptors and Percentages					Total
Extremely concerned	Moderately concerned	Somewhat concerned	Slightly concerned	Not at all concerned	33
6.1%	21.2%	36.4%	30.3%	6.1%	100%

Table 10 depicts sophomores, juniors, and seniors' views on how these types of tests still represent a moderate (66.7%) or minor (24.2%) problem (for freshmen (see table 7). However, BBTs seem more important for this population since 66.7% consider these tests a moderate problem, compared to 32.8% of freshmen. Comments such as "it seems that depending on how much you talk, may increase the mistakes." Your grade can change according to the topic or classmates" (P-76), or "it is not easy to talk in front of 3 professors about something I did not prepare" (P-94), attest to this perception.

Table 10. Sophomores, Juniors, and Seniors' Degree of Concern about the Test

Descriptors and Percentages					Total
A serious problem	A moderate problem	A minor problem	Not a problem at all	No Answer	33
6.1%	66.7%	24.2%	0%	3.0%	100%

For a direct comparison of foreign language anxiety levels, results from tables 8, 9, and 10 (sophomores, juniors, and seniors) can be contrasted with those from tables 5, 6, and 7 (freshmen).

The Test and Professors' Instructional Choices

For freshmen, a consensus seems to exist on the compatibility between the BBT and their instructors' methodological preferences during the course. When asked whether the exam format reflected the methods used by the professors (see table 11), 41.8% stated that they agreed and 35.8% reported to agree strongly with the statement.

Table 11. Freshmen's Agreement on Compatibility of Test Format and Procedures with Instructor's Methodology

Descriptors and Percentages						Total
Strongly agree	Agree	Neither or N/A	Disagree	Strongly disagree	No Answer	67
35.8%	41.8%	9%	4.5%	1.5%	7.5%	100%

On the question of whether the test directly influenced the classroom activities chosen by the instructors (see table 12), 34.3% of the informants expressed that for the most part their professor's choices seemed to be influenced by the BBT, while only 4.5% thought otherwise. From a qualitative perspective, the results appear slightly more varied. Some test-takers expressed moderate criticism through comments such as "the instructor could use a better methodology" (P-39, sic), "I think more speaking and oral expression or reading during class would've been better" (P-14, sic), "listening and speaking [...] worked, but nor evaluated with judges taking notes" (P-65, sic), and "I would ask for a bit more of activities like the one we had on the last Friday (impromptu) but the procedures were good" (P-48, sic). Others stressed full agreement: "yes they [activities] did [match the procedures;] we got strongly prepared in class for taking this test" (P-50, sic); and yet others stated direct agreement but did not answer how the test impacted instructional choices: "I agree with all the procedures" (P-40).

Table 12. Freshmen: Influence of Test on Classroom Activities

Descriptors and Percentages						Total
Always true	Sometimes true	Neutral	Rarely true	Never true	No Answer	67
20.9%	34.3%	28.4%	4.5%	4.5%	7.5%	100%

Quantitative data shows that for sophomore, junior, and senior test-takers, the consensus on whether or not the BBT is compatible with the methodology used by the professor seems slightly less clear-cut than for freshman students. Although 33.3% agree on the congruence between test format and classroom methodology, only 3.0% fully agreed strongly with it. Table 13 summarizes these results.

Table 13. Sophomore, Junior, and Senior's Agreement on Compatibility of Test Format and Procedures with Instructor's Methodology

Descriptors and Percentages						Total
Strongly agree	Agree	Neither or N/A	Disagree	Strongly disagree	No Answer	33
3.0%	33.3%	27.3%	24.2%	9.1%	3%	100%

The previous consensus remains true for whether or not the BBT influenced the classroom activities chosen by the professor. In this sense, 36.4% of the informants agreed that the test sometimes influenced the instructors' choice of classroom activities, 27.3% remained neutral, and another 27.3% assessed this influence as being *rarely true*, as shown in table 14.

Table 14. Sophomores, Juniors, and Seniors: Influence of Test on Classroom Activities

Descriptors and Percentages						Total
Always true	Sometimes true	Neutral	Rarely true	Never true	No Answer	33
6.1%	36.4%	27.3%	27.3%	0.0%	3%	100%

On the qualitative side, general comments lean toward a negative perception of this relationship. For example, P-77's major criticism is that "the activities made in the classroom in big groups [are too different from] two people talking in front of three professors", and that "sometimes the activities made in the course were not similar to what they called 'tribunal'" (P-77). Along the same lines, P-85 argues that even though they receive a good deal of speaking practice for the BBT, "it is not what is done throughout the course" (P-85), just like P-100, who claims that "in some cases the professor paid close attention, but in others, it was just like a reading class or just watching videos" (P-100, sic). Lastly, P-80 criticizes the gap between professors' language proficiency and lack of assessment literacy: "Some professors are extremely proficient, but they are somehow limited when it is time to test students" (P-80). Taken together, the two groups report somewhat similar opinions in terms of the consistency between the BBT and the professors' methodological choices, but they differ in their qualitative assessments of this relation (see *Discussion* section).

The Test and Professors' Decision Making

The last variable explored whether students felt that the BBT was a good basis for professors' decision-making for planning, teaching, and assessment in future courses they would teach. Freshmen were somewhat optimistic: 34.3% predicted likelihood and 32.8%

foresaw extreme likelihood that the test results would be used for such purposes, as shown in table 15.

Table 15. Freshmen: Instructor's Likelihood to Use BBTs for Future Decision Making

Descriptors and Percentages						Total
Extremely likely	Likely	Neutral	Unlikely	Extremely unlikely	No Answer	67
32.8%	34.3%	20.9%	3%	1.5%	7.5%	100%

The following comments offer the qualitative counterpart for the numbers above:

> "I hope so!" (P-1)
> "They shouldn't, so I think they won't, but at the end it's their choice, so they might consider it." (P-14, sic)
> "I am sure it will be like that, they always seemed to be worried about their students, so, I am sure they will implement our feedback in a positive way." (P-38, sic)

Table 16 shows that for sophomores, juniors, and seniors, the appraisal of this variable seems slightly less optimistic, with only 12.1% predicting extreme likelihood and 45.5% stating likelihood. The comments, however, showed divergent opinions on this matter. P-80 stated: "From a professor's perspective, I consider that results of a test like this shouldn't influence a decision regarding future courses to teach" (sic); P-84 claims that "every student is different to take the results as absolute would be completely wrong." (sic); and P-100 criticizes some professors' unlikelihood of changing their teaching practices based on BBT results: "I know some of them do take it into account. However, after taking some courses with the same professor or listening to others' comments, I know some do not change anything about the way they teach" (P-100).

Table 16. Sophomores, Juniors, and Seniors: Instructor's Likelihood to Use BBTs for Future Decision Making

Descriptors and Percentages						Total
Extremely likely	Likely	Neutral	Unlikely	Extremely unlikely	No Answer	67
12.1%	45.5%	18.2%	18.2%	3%	3%	100%

Concluding Remarks

Discussion of Findings

This paper has aimed to explore how much board-based oral tests influence language proficiency, foreign language anxiety levels, instructional choices, and decision making in three ELCL-UNA three bachelor's level programs. Our discussion begins with the first area investigated: the BBT and students' language proficiency and foreign language anxiety levels. Regarding the increase of students' language proficiency, findings suggest a correlation between informants' own proficiency level and their perceptions about such correlation; that is, the higher their English level, the less of a connection they perceived between these two elements, just as Reynolds[84] had found in a study of students' perceptions of the washback of the TOEFL test. On the relation between the test and students' foreign language anxiety levels, results show high levels of emotional distress mainly before the BBT, even amongst students who had been in the program for longer periods of time. To the uncritical eye, this finding is contrary to expectations since more experienced learners may be expected to have become used to board-based testing and the dynamics it entails; however, this can also be evidence that summative assessment of this kind is more complex than we generally realize and that foreign language anxiety-generating factors are not necessarily correlative

84 Reynolds, iv.

to test-taking experience. Here, one must agree with Teemant's[85] assertion that more research is necessary to unveil the degree of importance that certain factors (foreign language anxiety included) exert upon test-takers' performance.

Turning now to the second area investigated (students' perceptions about professors' instructional choices and decision making), our analysis indicates two main tendencies. First, the two groups of informants (a. freshman students and b. sophomore, junior, and senior students) have seemingly analogous opinions in terms of the consistency between the BBT and the professors' methodological choices, but they differ in their qualitative assessments of this relation. For instance, some of them expressed that the BBT was not look at all like the activities used in class by the professor. That violates the principle of transparency in language assessment; as Rogier explains: "Transparency makes students part of the testing process by ensuring that they understand what the course objectives are and what will be tested, as well as the format of tests and how they will be used and graded."[86] Second, results were contrary to expectations in terms of the BBT and professors' decision-making. For example, most freshman students said they were positive that professors would use test results to make decisions about future courses they would teach, while the percentage of sophomores, juniors, and seniors who reported the same was much lower. Qualitative data from these students suggests that such views are due to previous experiences, with reports that some professors never update their teaching methods. While our study did not probe the reasons for these opinions, we are left with the question of whether this happens due to lack of willpower (if it actually happens), or due to lack of training, as Chen reported in her 2002 study. According to this author,

85 Teemant, 101.

86 Dawn Rogier, "Assessment Literacy: Building a Base for Better Teaching and Learning," *English Teaching Forum* 3 (2014): 2-13. Available at: <https://americanenglish.state.gov/files/ae/resource_files/etf_52_3_02-13.pdf>.

educational reform (such as the officializing of a board-based testing system) tends to be superficial for teachers since the content of their teaching is affected, but their methodologies remain unchanged due to lack of guidelines on how to align policy reforms with instruction.[87] At any rate, for now these questions are outside the scope of our research.

Despite the evidence analyzed so far, it is unclear whether higher-level students have more negative views toward the BBT because they have the language skills to express it, because they have been part of the program for a longer period of time, or because of aspects not yet clear in this study.

Implications, Limitations, and Further Research

Through an analysis of BBTs and students' perceived proficiency levels and foreign language anxiety, as well as their views on the BBTs and professors' instructional choices and decision making, our study has provided baseline evidence to understand the connections between this test and the four factors mentioned above. For the field of second language assessment, the inquiry has uncovered preliminary perspectives on the issue and opened an avenue for future research on the sub-field. It has generated insights for curricular decision making and planning, especially in the light of the UNA accreditation processes since 2006.

For more conclusive results, however, future research should address a number of limitations faced by the current research. The first and most obvious is that the professors' perspectives are yet to be investigated, to complete the spectrum of opinions about the issue at stake. Second, larger student and instructor samples, as well as wider contexts should be included as a way to test our results and address generalizations on the findings. Third, other methodologies such as narrative or phenomenological research could be incorporated to delve more deeply into the phenomenon under investigation.

87 Chen, 17.

Lastly, a broader range of educational actors should also be considered, including stake holders, policy makers and faculty from the ELCL French and Spanish Departments.

In the literature review above, we presented the following observation by J. Charles Alderson: "[In the field of second language testing,] the apparent progress we think we have made – that we celebrate at conferences and seminars like this one, that we publish and publicise – may well not represent progress so much as activity, sometimes in decreasing circles."[88] Almost thirty years later, we find Alderson's criticism quite valid. Our hope is that this study of the effects of board-based speaking tests can open an avenue of reflection, improved teaching, and future studies to help turn the decaying testing activity that Alderson feared into actual progress in the language education landscape.

88 Alderson, 24.

Appendix 1: Likert Scale of Agreement

Objective: This instrument gathers data on the students' perceptions about the impact of board-based oral tests on various factors of language instruction and learning.

General Instructions:

- Read the statements below.
- Choose the answer that best represents your opinion about the statement.
- Feel free to write comments on the spaces provided.

The information you provide will be used for research purposes only, and your identity will be kept confidential. Your honesty will contribute to the validity of the information collected.

Part A

1. The board-based (*tribunal*) test has helped increase my general speaking skills in English.

5 Strongly Agree	4 Agree	3 Neither or N/A	2 Disagree	1 Strongly Disagree

Comments: _____

2. The course contents studied in class helped me develop _____ speaking skills.

5 excellent	4 very good	3 good	2 fair	1 poor

Comments: _____

Part B

1. Before taking this test, I was _____.

5 extremely concerned	4 moderately concerned	3 somewhat concerned	2 slightly concerned	1 not at all concerned

Comments: _____

1. After taking this kind of tests, I am usually _____.

5 extremely concerned	4 moderately concerned	3 somewhat concerned	2 slightly concerned	1 not at all concerned

Comments: _____

2. In general, this test represents _____ for me.

5 a serious problem	4 a moderate problem	3 a minor problem	2 no problem at all

Comments: _____

Part C

1. The test format and procedures match the methodology used by the instructor in the course.

5 Strongly Agree	4 Agree	3 Neither or N/A	2 Disagree	1 Strongly Disagree

Comments: _____

2. I feel this test influenced the instructor' choice of classroom activities.

7 Always true	6 Usually true	5 Sometimes true	4 Neutral	3 Infrequently true	2 Rarely true	1 Never true

Comments: _____

Part D

1. I think the instructor will consider the results of this test to make decisions about future courses s/he teaches (Likelihood: Extremely unlikely, etc.).

5 Extremely likely	4 Likely	3 Neutral	2 Unlikely	1 Extremely unlikely

Comments: _____