Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies, 15(4), 1252-1268; 2019

Washback effect of Turkish foreign language proficiency exam YDS: A scale development study

Murat Polat [a]*

[a]*Anadolu University, School of Foreign Languages, Eskişehir 26470, Turkey*

APA Citation:

Polat, M. (2019). Washback effect of Turkish foreign language proficiency exam YDS: A scale development study. *Journal of Language and Linguistic Studies*, *15*(4), 1252-1268. Doi: 10.17263/jlls.668404

Submission Date:19/08/2019

Acceptance Date:28/09/2019

Abstract

Investigating the possible washback effect of Foreign Language Proficiency Exam called YDS could provide valuable data regarding how test-takers feel about taking this exam. This study aims to design a scale and collect data regarding the possible washback effect of YDS. In the initial phase, with the help of 6 academicians from ELT and testing disciplines, a preliminary draft of an attitude scale was developed. Secondly, the scale was piloted in order to carry out exploratory and confirmatory factor analyses. Thirdly, the newly developed scale's psychometric qualities (validity and reliability) were measured. After it was confirmed that the scale which had 4 factors and 13 items was a valid and reliable tool, the data including the views of Turkish academicians were gathered focusing on the washback effect of the YDS via this scale in the final phase. The overall data of this study were collected from 84 state universities with the contribution of 2683 academicians from different disciplines. The findings revealed that YDS has negative backwash effect on Turkish academicians in terms of certain sub-categories including test strategy, motivation, perception and test taker's psychology.

Keywords: Washback effect; foreign language testing; high stake tests; YDS; proficiency tests; test impact; exploratory factor analysis; confirmatory factor analysis

1. Introduction

The terms "washback" or "backwash" have been commonly used to refer to the impact of any kind of testing not only on teaching and learning but also on materials and curriculum development process. Research in testing (Alderson, 1986; Linn, 2000; Wall, 2000) has revealed that a single test score could not be considered a numeric indicator of a particular ability which is tested since test scores are significantly affected by many variables including the type of the test, its content, test takers' behaviors, strategies applied by the test takers and a number of inferences which could be made on those scores. Thus, it could be possible to infer that tests and their consequences might affect not only the test takers' academic careers but also the educational strategies and policies of the institutions and countries as well. Similarly, Shohamy (2006) and Wall (1996) have underlined the indirect effects of high-stakes tests and

* Corresponding author. Tel.: +90-222-335-0580
E-mail address: mpolat@anadolu.edu.tr

stated that such tests have a great impact on the educational policies of the society since those tests are sometimes used to manipulate or implement certain educational plans. Considering those possible effects, "test washback" studies could provide critical information to education policy makers, decision makers and test writers about the decisions they make and let them gain deeper insights on either positive or negative effects of the tests they design and administer.

2. Literature review

Testing or assessment in a broader sense is mostly considered as the final step of an instructional process which might reveal the results of it and could lead to some revisions or changes in the curriculum, materials, teaching hours, etc. (Airasian, 1994; Brown, 1998; Heaton, 1988). However, from the view of students, testing might have far more denotations than we used to infer (Linn, 2000). Test types and their contents demonstrate what is valued, what is going to be taught and surely, what is going to be assessed (McEwen, 1995). Therefore; students, by examining the previous tests, mostly prioritize how they are going to be tested and mostly make their own educational plans rather than the plans launched by education planners or their institutions. In order to be successful, they value what test makers value and this could be considered the main impact of such tests. Pearson (1988) supported the idea that the intended direction in which backwash effect leads must be forward, which could be interpreted in two ways. First, it could be explained as the positive effect of a test on students' development and achieving the intended educational outcomes. On the other hand, it could be interpreted as the leading role of testing on learning and teaching process (Cheng & Curtis, 2004). Thus, a good test is supposed to create positive effects on learning whereas it is also possible that tests could sometimes have undesirable (negative) effects on students' learning.

Alderson and Wall (1993) underlined the undesirable effects of tests and criticized this negative influence of exams on students' learning and their self-development. They stated that because of standardized tests, teachers and students end up teaching and learning towards the format of the tests, ignoring the nature of real education for lifelong needs. Green (2007) also highlighted the importance of washback effect and believed that if a test does not support students' learning but causes anxiety and reaction to further studying, the backwash effect is negative and undesirable. Özmen (2011) and Ching Pan (2009) have also studied positive and negative effects of tests, and they stated that those effects may operate at macro (students' personal understanding in terms of education, educational policies and decision-making mechanisms) and micro (classroom practices, study habits, etc.) levels.

In another study, washback effect of a test in a school was observed in a quite different way. Fish (1988) revealed that high school teachers admitted that they felt nervous when they display classroom scores publicly. He also added that less experienced or novice teachers have this anxiety and accountability stress more when they announce test scores. This problem was reflected in Fish's study at micro level; however, the effects of high-stakes tests and their results could be much worse at macro level. In their study, Noble and Smith (1994) underlined a very important fact about standardized tests and they revealed that high-stakes tests could affect educational programs and teachers' way of teaching directly and negatively. They found that teachers' strategy training on test-taking skills and assisting the learners in finding the correct choice on multiple-choice test worksheets might lead to better scores in the short term but might spoil the nature of meaningful and deeper learning in the long run. In a similar study, Smith (1991) announced some detrimental effects of high-stakes testing on learning as those tests change how students learn, reduce the lecture time for effective instruction and most importantly minimize teachers' creativity and teaching ability to lecture main concerns more and to use pedagogic methods and classroom materials which are not suitable for multiple-choice test formats.

In the Turkish context, washback effects of high-stakes tests were reported (Akın, 2016; Cinkara & Tosun, 2017; Hatipoğlu, 2016; Külekçi, 2016; Özmen, 2011; Yavuzer & Göver, 2012; Yeşilyurt, 2016) to be negative since most teachers at state schools and even at private schools feel confined by the limitations of multiple-choice test contents, and similarly students mostly prefer to focus on how they are going to be tested, what could be asked and how they would respond to those questions. Of those high-stakes tests, YDS (Foreign Language Proficiency Exam) which has been one of the major foreign language proficiency tests in Turkey since 2013 is the main concern of this study. YDS is a multiple-choice test made up of 80 questions which are mostly calibrated to measure students' grammar, vocabulary, reading and translation skills in a particular foreign language. YDS which does not contain speaking, writing and listening sections has been criticized by many researchers (Külekçi, 2016; Polat, 2018; Yavuzer & Göver, 2012; Yeşilyurt, 2016) in terms of its negative backwash effects not only on candidates, but also on language learning methods, materials and teaching techniques in Turkey.

Considering the students' language learning habits, Külekçi (2016) stated that YDS has a negative washback effect since learners spend more time on practicing test techniques rather than studying and learning the target language. Checking the previous exam questions and employing some grammatical formulas seemed more beneficial for the test takers in order to respond more strategically to tricky grammar and vocabulary questions. Similarly, Polat (2018) criticized the effect of YDS on academicians and supported the idea that such multiple-choice foreign language exams value and make learners value the mechanical and grammatical side of a foreign language rather than its communicative sides. He added that since taking a critical score like 65-70 points would enable the academicians have their academic titles, to be promoted or to be accepted to M.A. or PhD. programs, university teachers in Turkey mostly overlook the development of some communicative skills such as fluent speaking or writing in the target language while studying that particular foreign language. Finally, Yavuzer and Göver (2012) stated that because of such high-stakes language tests in Turkey, students and academicians spend their valuable time studying explicit grammar rules and academic vocabulary rather than focusing on their academic/scientific studies. Memorizing some specific advanced vocabulary items, studying long lists of grammatical rules and disregarding the real goal of learning a foreign language helps nobody, neither the test designers nor the policy makers.

Considering the findings of the related studies on washback effects of language tests in Turkey, it was clear that it could be useful to develop a scale to study this phenomenon, measure its findings statistically and collect data from university teachers who could provide valuable data regarding the impacts of such tests. The results and their reflections could shed a light on foreign language education in Turkey and might reveal some important aspects of language education including the ways we teach, practice and test. It might also lead teachers, administrators and policy makers reconsider the definition of foreign language competency and review proficiency assessment applications in Turkey.

2.1. Research questions

Since this study aims to develop a scale for measuring test washback effect, it has two major goals. To begin with, it was aimed to develop a new, reliable data collection tool to investigate the academicians' attitudes towards language testing and YDS in Turkey. Secondly, Turkish academicians' attitudes towards the language test YDS were examined according to several independent variables. The study tries to answer the following research questions:

1. What are the results of the exploratory factor analysis of Academicians' Foreign Language Test Attitude Scale (AFTAS)?

2. What are the results of the confirmatory factor analysis of Academicians' Foreign Language Test Attitude Scale (AFTAS)?

3. Considering the accepted reliability degrees, are the test results of Academicians' Foreign Language Test Attitude Scale (AFTAS) substantial?

4. Are there significant differences among academicians' attitudes towards YDS according to their genders, disciplines, titles and their aims in taking this exam?

3. Method

In order to be able to explore academicians' attitudes on YDS, there were two major aims in this research. First, it was planned to construct a new and reliable scale to gather data on academicians' beliefs about the foreign language test YDS, next it was intended to identify Turkish academicians' feelings and attitudes towards YDS and to test if their attitudes change according to some independent variables including gender, age, discipline, title and participants' aims in taking this exam.

3.1. Participants

As for the research model, the general screening model was applied and convenience sampling technique was preferred to collect data. AFTAS was designed with the help of 296 voluntary academicians working in 4 state universities in Eskişehir and Ankara. Those 296 participants were randomly divided into two groups to define and check the factor structure of the scale. Therefore, the first group's (149 participants) responses were used to identify AFTAS's qualitative properties and the emphasis was on the scale's construct validity, reliability and Exploratory Factor Analysis' results. After defining the factor structure and making necessary revisions (the scale had 19 items in the initial phase and after the exploratory factor analysis this number was reduced to 13 because 6 items were deleted), the second group's (147 participants) responses were used to find out the modified scale's internal validity level using Cronbach Alpha coefficient and Confirmatory Factor Analysis results. The data gathered form the scale development procedure were not used in defining academicians' attitudes against YDS since the aim in the initial data collection phase was to pilot the scale's first draft and its modified version.

After the scale development phase, the survey questions were sent via e-mail and regular post to all state universities in Turkey, and a total of 2697 academicians working in 84 universities participated in this study. Of those participants, 14 were excluded from the study since seven academicians did not provide some demographic data and seven did not complete one or some parts of the Likert scale which was developed to collect data. As a result, 2683 valid data forms were used in this research. Considering the genders of the participants, a balanced gender distribution was observed. 51.7% of the overall participants were females and 48.3% were males. In terms of age distribution, there was again a balanced distribution. 16% of the participants were between 22-27, 32.4% of the participants were between 28-33, 25.2% of the participants were between 34-39 and the final group participants who were over 40 were 26.5%. As for the titles of the participants, the majority of the participants were research assistants and lecturers (1712 participants), and considering their disciplines there was again a balanced distribution among social sciences (25.4%), educational sciences (32.8%), Master of Science (27.3%) and others (14.4%). Finally, in terms of their aims in taking this exam, participants who take this exam to be promoted and to be accepted to Master's and PhD. programs formed up the majority of the participants (70.1%).

3.2. Instrument(s)

Development of a reliable and valid scale in social sciences is not a simple process and requires an extensive study which takes into consideration the number of items to be used, focus participant group, time allocated and some sociological variables such as the context, culture, etc. Sometimes it might be observed in similar studies that expert check and confirmation are thought to be enough for a scale to be used in data collection; however, some critical steps should be followed in the generation of a new scale. Although it is very difficult to reach a consensus on the development process, researchers in the field suggested that a valid, reliable and practical scale can be developed by following a certain number of steps (Büyüköztürk, 2013; Crocker & Algina, 1986; Seker & Gencdogan, 2014, Vieira, 2011). The following steps listed below were suggested and followed to develop a (five-point) Likert Scale called Academicians' Foreign Language Test Attitude Scale (AFTAS):

(1) Defining the main objective, participant group and the required time.

(2) Outlining the focus and defining the objective of the scale.

(3) Item writing for the draft items considering the scope of the scale.

(4) Expert check, revision on items and inclusions/exclusions.

(5) Determining data collection method and procedure, data analysis.

(6) Piloting the scale with a specific group (Exploratory Group).

(7) Testing the factor structure with another group (Confirmatory Group).

(8) Comparison of the results and analyzing the data.

3.3. Data collection procedures

Before an attempt to design a scale for data collection, an extensive literature review was implemented in order to check if there were other scales that could be used related to the focus of this research. The literature review revealed no valid and reliable scale that was developed to collect academician's opinions regarding YDS practice in Turkey. Therefore, it was planned to create a new scale, and firstly the objective and the intended outcomes of this research were reported to the people who were in charge in academic issues working in Turkish state universities to announce their faculties to contribute to this study. Then, within the first participant group, 296 academicians working in 4 different universities were set apart from the main participant population to develop the scale. It was assumed that the required time for the whole study could be around 12 months. In the next step, considering the previous studies in the literature and the criticisms academicians made on YDS, 40 items were written to test participants' attitudes towards this language test. After a brief review with two testing experts from a state university in Eskişehir in terms of the wording of the items, the number was reduced to 28 items. In the next step, 4 experts including 2 professors and two assistant professors from ELT departments reviewed the scale, and they agreed that 9 items could be excluded since those items in some ways were overlapping with others or inquiring the reliability of YDS indirectly. In the next step, the preliminary data set from the Exploratory Group was collected and examined statistically. After making the necessary revisions on the scale (the number of items was reduced to 13). 2 months later, Confirmatory Group's data were collected and analyzed statistically in order to finalize the scale and to be able to control and approve the initial factor loads which were found in the data set of the Exploratory Group. When the scale was proven to be valid and reliable for data collection, it was sent to 84 state universities in Turkey to collect as much data as possible on academicians' attitudes towards YDS. This process took another 4 months and the data collected from 2683 (from the total 2697, 14 were excluded) participants from different universities were analyzed and reported.

3.4. Data analysis

The first step in the development the new scale, Kaiser–Meyer–Olkin (KMO) and Bartlett Sphericity tests were conducted to be able to measure the relevance of applying factor analysis on the set. Later, Varimax rotation, correlation of anti-image, Cronbach Alpha for the reliability, exploratory and confirmatory factor analysis to be able to find out the factor structure and test and re-test the factor loads of AFTAS were conducted. IBM-SPSS 21 and IBM-AMOS programs were used in these analyses. In the third phase, a number of statistical analyses were planned to be utilized by the application of t-test, ANOVA and Tukey Post Hoc Tests to test if the attitudes of academicians towards YDS differ considering some variables such as gender, discipline, titles and their aims in taking this exam. Having taken the advice of the experts in the field, before utilizing the statistics listed above, Kolmogorov-Simirnov with Shapiro-Wilk and Homogeneity tests were applied to check the distribution (in terms of normality and homogeneity) of the data. The results revealed that the data had a normal distribution considering the results of both Kolmogorov-Smirnov and Shapiro-Wilk tests ($p>0.05$). Thus, parametric tests were utilized to compare the groups; however, in the homogeneity test, it was found that the data were not homogeneous. Normally ANOVA is preferred (Green & Salkind, 2008; Siegel, 1977) when two or more groups' means are compared, but when it is desired to compare the variances of those groups' means, they should be equal to run parametric tests like Tukey. However, since the homogeneity among groups was not equal, Tamhane test was preferred to compare the variance among the groups.

4. Results

4.1. Exploratory Factor Analysis of AFTAS

In most scale developments studies, there are certain steps to follow (Büyüköztürk, 2013; Doğan & Doğan, 2014; Özdamar, 2013) such as checking if the data are available for factor analysis, the number of sub-dimensions and the items' factor loads and correlations. That's why, the data collected from the first group academicians were analyzed as the first step to see if they were appropriate to factor analysis or not. Principal Component Analysis method was used to check the construct validity of AFTAS. First, in order to identify whether the data were appropriate for factor analysis, KMO (Kaiser–Meyer–Olkin) Test and Bartlett Sphericity Tests were both conducted as the initial phase of principal component analysis. To have a more vivid view of the factor structure, varimax rotation method was used on the data gathered from academicians. Kaiser–Meyer–Olkin value calculated for the data set was 0.709. Any value of KMO higher than the critical limit 0.50 shows that the data set is appropriate for running factor analysis. As for the Bartlett Test, the result was 0,0001 [χ^2 = 516,800; df=148, $p<0.01$]. The data sets' significance was found to be smaller than 0.01 and it implies that factor analysis can be made on the set. Next, principal component analysis revealed that the items 7, 14, 15 and 16 gained lower factor loads than the expected value (0.300) within the total item correlation amounts. In addition, the items 10 and 12 formed up a single factor together; however, to have a sound factor combination, there should be at least 3 or more items in a factor that could be explained by the scale, and those items should have high factor loads (Büyüköztürk, 2013; Özdamar, 2013).

Finally, the items, 7, 10, 12, 14, 15, and 16, totally six items were taken out of the new scale, AFTAS. In Table 1, it could be checked that the estimated factor loads of the included 13 items differ from 0.392 - 0.703 and item total correlation amounts range from 0.361 - 0.762. With the help of varimax rotation, the variance of the 4 factors reveals the attitudes of academicians towards YDS up to 59.609% in total and the Cronbach Alpha was 0.743 which could be classified as a reliable scale. Item factor loads and total correlations are given in Table 1.

Table 1. AFTAS Factor analysis

Item Number	Initial Factor Load	Item Total Correlation	Item Number	Initial Factor Load	Item Total Correlation
1	0.513	0.469	**9**	0.392	0.361
2	0.435	0.396	**11**	0.422	0.479
3	0.581	0.623	**13**	0.598	0.642
4	0.402	0.398	**17**	0.703	0.762
5	0.472	0.549	**18**	0.448	0.512
6	0.509	0.611	**19**	0.467	0.536
8	0.483	0.563			
Variance four factors explain = **59,609%**			Cronbach Alpha = **0.743**		

It is stated that (Büyüköztürk, 2013; Özdamar, 2013) values over 0.300 for initial factor load and item total correlation could be accepted in exploratory factor analysis. The values presented in Table 1 are all above the critical limit 0.300, which means that they are acceptable. Moreover, the entire variance that might be revealed by this model was found as 59%, which is another reasonable value for scale development (Doğan & Doğan, 2014). Table 2 displays the anti-image correlation degrees of the items.

Table 2. AFTAS Anti–image correlation degrees

Item No	Anti-image	Item No	Anti-image
1	0.816	**9**	0.863
2	0.721	**11**	0.895
3	0.840	**13**	0.907
4	0.713	**17**	0.958
5	0.782	**18**	0.863
6	0.842	**19**	0.878
8	0.902		

Anti-image screening is a useful method to check the factor loads and see if they contribute to the factor structure, and the values over 0.50 present that such items contribute to the sample model. As the item list in Table 2 reveals, all the existing items anti-image values are over the critical limit, they range between 0.713 to 0.958, which means that all the items significantly contribute to the factor structure. Next step is to find out the sub-categories driven from this structure. Many researchers (Büyüköztürk, 2013; Doğan & Doğan, 2014; Özdamar, 2013) agreed that varimax rotation is an effective method in clarifying a factor structure and simplifying the expression of a single factor besides a number of some other major factors. Table 3 reveals the results of varimax rotation.

Table 3. AFTAS Rotated component matrix

	Component			
	1	**2**	**3**	**4**
Item 5	.741			
Item 4	.729			
Item 2	.712			
Item 11	.651			
Item 9		.754		
Item 13		.730		
Item 3		.698		
Item 6			.731	

Item 1	.712	
Item 19	.642	
Item 8		.869
Item 17		.673
Item 18		.606

It could be seen from the results in Table 3 that there are 4 main subcategories in the scale. After the varimax rotation, for the first subcategory items, 2, 4, 5, 11 could be listed, and all these items were related to test strategies; therefore, this factor was named as “Test Strategy” and the listed items were re-coded as 1, 5, 9 and 13 respectively. For the second subcategory, items 3, 9 and 13 were listed, and these were related to test taker’s motivation, so the second factor was named as “Motivation” and the listed items were re-coded as 2, 6 and 10. For the next subcategory, items 1, 6 and 19 were listed and these were related to test taker’s exam perceptions, so the third factor was named as “Perception” and the listed items were re-coded as 3, 7 and 11. For the last subcategory, items 8, 17 and 18 were listed and these were related to the test taker’s psychology, so the fourth factor was named as “Test taker’s psychology” and the listed items were re-coded as 4, 8 and 12.

4.2. Confirmatory Factor Analysis of AFTAS

Final step of creating the new scale was to confirm its factor structure and check if the subcategories driven from exploratory analysis would be verified by confirmatory factor analysis. The estimated model of “AFTAS” is presented in Figure 1.

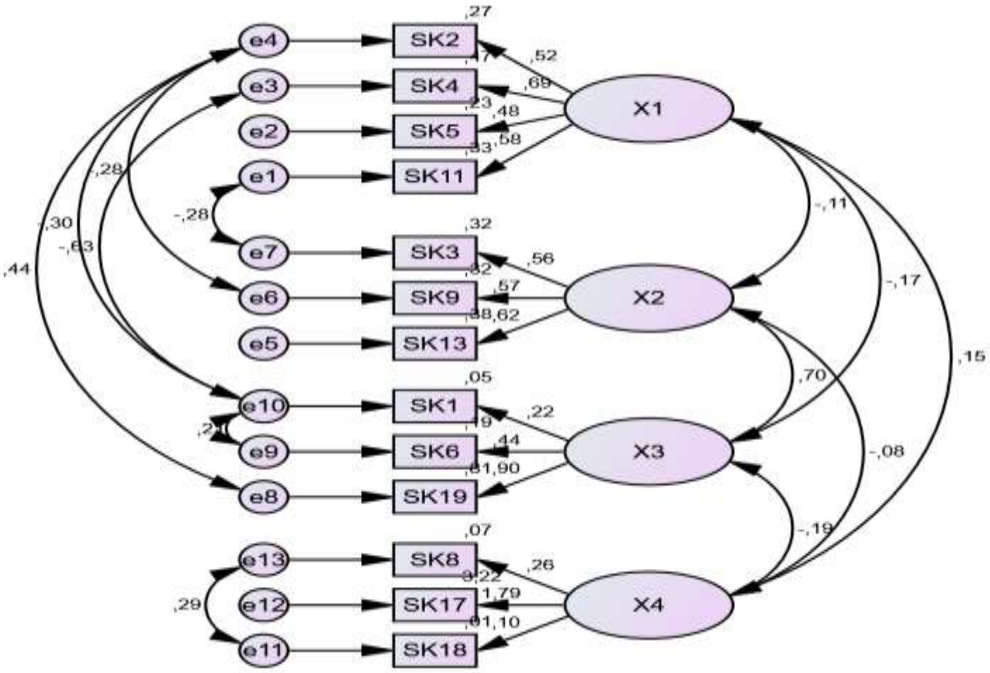

Figure 1. AFTAS Confirmatory factor analysis model

It could be checked that Confirmatory Factor Analysis (CFA) verified the factor loads of AFTAS with four sub-categories and the correspondent items. The results of the analysis also revealed the Chi-square and degree of freedom values (χ^2 =106.524, (df=52, p<.01), (χ^2 /df=2.049)). Kline (2005) and Vieira (2011) reported that if the df value is below the critical 3.0 value, this could be assumed as a perfect match. Finally, the RMSEA (root mean square error of approximation), AGFI (Adjusted

Goodness of Fit Index) and RMR (Root mean square residual) values were checked to find out the goodness of fit. Table 4 reveals the results the CFA analysis in terms of goodness of fit.

Table 4. CFA Goodness of fit values

χ^2	D_f	χ^2/D_f	RMSEA	AGFI	RMR	CFI
106.524	52	2.049	0.073	0.859	0.040	0.920

RMSEA, which could be stated as the most critical value to check the goodness of fit, was calculated as 0.073, and it could mean a good match since it was lower than the critical value 0.08 (Doğan & Doğan, 2014; Kline, 2005; Vieira, 2011). Next, AGFI value was 0.859 which was above the critical value 0.080 and could be acceptable for the match. As for the RMR, if it is lower than 0.10 and it would mean a good match (RMR was 0.040). Furthermore, CFI which should be more than 0.90 was calculated as 0.92 which was again an evidence of a very good match between the data and the acquired model (Kline, 2005; Vieira, 2011). The last but not the least, the reliability of the test was checked, and it was found reliable enough for data collection (Cronbach Alpha = 0.769). All in all, it could be reported that that the 4-factor structure of AFTAS was confirmed by the findings driven form CFA.

4.3. AFTAS Findings Related to Academicians' Attitudes Towards YDS

As it was mentioned before, 2683 Turkish academicians' responses were collected and analyzed through the newly developed scale AFTAS (Academicians' Foreign Language Test Attitude Scale). AFTAS was designed as a 5-point Likert scale changing from 1 (Strongly disagree), 3 (Partly agree) to 5 (Strongly agree). The scale had 4 sub-categories as "Test Strategy, Motivation, Perception and Psychological Effect"; therefore, results of each factor will be presented in this section separately under the scale's factor names.

Participants' responses to the first part of AFTAS, score means and their standard deviation values are shown in Table 5.

Table 5. AFTAS factor 1 test strategy item statistics

Item	Mean	Std. Dev.	N
1. YDS does not include listening questions that is why listening comprehension skills are disregarded in this test.	4.25	.775	2683
5. To be successful in YDS, a number of test techniques and strategies should be practiced.	3.95	.982	2683
9. Since YDS has advanced vocabulary questions, memorizing new words is a common way of language learning.	4.23	.856	2683
13. To have a tutor for YDS preparation can be a good idea.	3.92	.867	2683

The items of AFTAS related to foreign language test strategies revealed that participants have mostly negative attitudes towards YDS since the test mostly motivates testees to learn test strategies and memorize some vocabulary items rather than learning the foreign language. It should be reminded once more that the Likert scale had answer codes ranging from 1 (Strongly disagree) to 5 (Strongly agree) and when the responses were checked again it is clear that according to academicians' views, YDS is reported to be a complex foreign language proficiency test that requires strategy training, mentor help and memorizing words that would unlikely to help testees learn and use the target language in real communication. Another issue was to check if those responses differ according to the participants' genders, academic disciplines, titles and their aims in taking this exam. Considering the 1st Factor of AFTAS (Test Strategy), in terms of gender and academic discipline (studying in educational or social sciences, etc.), there was no significant difference among academicians' responses (p>0.5). However, when the other independent variables like participants' academic titles and aims in taking this test were considered, significant differences among responses were observed (p<.05). Academicians who took this exam to have extra payment from government or the ones who have professor or associate professor titles responded more leniently compared to the other academicians. Considering the fact that professors and associate professors have no obligation to present official proof for their foreign language competency levels, they reacted more positively to the effects of YDS since they do not need to be assessed by such exams any more.

Table 6. AFTAS factor 2 motivation item statistics

Item	Mean	Std. Dev.	N
2. YDS makes the test takers feel that learning a foreign language is easy.	1.07	.975	2683
6. YDS motivates the test takers to learn a foreign language.	1.38	.785	2683
10. YDS positively contributes to learning a foreign language.	1.23	.903	2683

Another important issue in test washback effect is to check and see how well a test can motivate the test takers. Considering the items of AFTAS under the motivation factor, it can be easily seen that participants disagree with the idea that YDS is a motivating test for academicians to learn a foreign language. Especially, the responses to Item 2 reveal that the test is highly difficult and does not contribute to participants' foreign language learning. Considering the 2nd Factor of AFTAS (Motivation), in terms of gender and academic discipline, there was again no significant difference among academicians' responses (p>0.5). However, when the other independent variables such as participants' academic titles and aims in taking this test were considered, significant differences among responses were observed (p<.05) in terms of motivation. Professors and associate professors believed that the test somehow makes people study and learn a foreign language, and they feel that this obligation directly or indirectly functions as a way motivation.

Table 7. AFTAS factor 3 perception item statistics

Item	Mean	Std. Dev.	N
3. YDS includes many grammar questions that lead people study grammar more while learning a foreign language.	4.16	.800	2683
7. YDS affects test takers negatively while preparing for the test.	3.98	.713	2683
11. Test takers who study for YDS could use the target language effectively.	1.49	.761	2683

People's perceptions about a test's negative or positive effects is another concern to figure out to what extent people trust that particular test's positive effect which would make the test a working one. Unfortunately, the items of AFTAS under the perception dimension showed that participants disagree with the idea that AFTAS has a positive effect on them and motivates the academicians to learn and use the target language more effectively. Especially, the responses to Item 3 reveal that the test leads test takers to study grammar rules more rather than to practice the language in speaking and writing contexts.

Unfortunately, this is a common problem in most developing countries. Not only learners but also teachers and decision makers do overvalue grammar and vocabulary knowledge and prioritize those language skills in language classes and language tests which lead students to prioritize those skills as well. Considering the 3rd Factor of AFTAS (Perception), in terms of gender and academic discipline similar to the other factors, there was no significant difference among academicians' responses (p>0.5). However, the variables such as participants' academic titles and aims in taking this test made significant differences among participants' responses (p<.05) in terms of their language learning perceptions. Especially, the responses to Item 11 prove the fact that studying for YDS does not contribute to the effective use of the target language.

Table 8. AFTAS factor 4 test takers' psychology item statistics

Item	Mean	Std. Dev.	N
4. Test takers are under stress before and after YDS.	4.11	.872	2683
8. YDS affects people negatively in terms of language learning.	3.81	.995	2683
12. YDS demotivates test takers.	3.58	.989	2683

Last but not least, another important effect of a test on test takers is its psychological effect. Determining if a test has positive or negative effects on people could give valuable data for decision makers since the psychological burden of an exam could demotivate learners and could possibly impede their efforts in language learning. When the test taker's psychology dimension was examined, it was found that the test causes stress and demotivates academicians in learning a foreign language. Considering the 4th Factor of AFTAS (Psychological Effect), in terms of gender and academic discipline there was no significant difference among academicians' responses (p>0.5). However, the variables such as participants' academic titles and aims in taking this test again made significant differences among participants' responses (p<.05) in terms of the test taker's psychology dimension. The responses to Item 4 reveal the fact that a group of academicians including the research assistants and assistant professors feel a great pressure while taking this exam since their promotions in academic ranking depend on the results of this test. Professors and lecturers which could be thought as the top and bottom academic ranks in Turkey have slightly more positive attitudes against the test since they do not have much expectation from the test results compared to other academicians.

5. Discussion

Determining test impact and taking its results into account can be highly beneficial for test designers, decision makers and teachers since no test could be evaluated independently out of its educational context. Thus, gathering information on how test takers feel about taking language tests, what learning habits those tests develop (directly/indirectly) and what implications could be made from the responses are valuable. For this reason, research on washback effect of tests could help decision makers see if they are on the right track, if the test outcomes are satisfactory or if the test takers are motivated to learn and practice more to be able to achieve their educational goals. Keeping this critical role of washback research in mind, this study was intended to investigate the washback effect of YDS (Foreign Language Proficiency Exam) which is a language proficiency exam in Turkey. Academicians and university students who want to enroll in universities' M.A. or PhD. programs in Turkey mostly take YDS to document their foreign language proficiency levels. The test does not include writing, listening or speaking sections; therefore, researchers and test designers are doubtful if it is a valid language test considering its deficiency in testing productive skills in the target language. Moreover, having questions mostly on grammar rules, complex reading passages, advanced vocabulary items and translation skills from one language to another, YDS is also criticized for having negative washback effect on language learners since today's mostly embraced communicative skills hardly involve some of those. Who would practice speaking, listening or writing in the target language if none were measured and considered as vital skills in the test? On the contrary, memorizing words and grammar rules are prioritized by Turkish language learners just because of the content and format of YDS apart from what normal language learners do throughout the world.

This study revealed that findings related to the washback effect of YDS were mostly negative and the responses of academicians were significantly different depending on their academic titles and their aims in taking this foreign language test, but the responses did not show any significant difference considering the participants' genders and academic disciplines. In terms of their academic titles, participants' responses were observed to be in 2 main groups. The first group included lecturers (who teach foreign languages or the academicians who have little or no intention to be promoted in the future), professors and associate professors (in Turkish context professors and associate professors do not need to document their language proficiency levels after they get those titles). On the other hand, research assistants and assistant professors (in Turkey those academicians must prove their foreign language proficiency in order to get a promotion) made up the second group. Considering the overall responses, academicians in the first group were slightly more positive (Mean= 2.08) towards the washback effect of YDS compared to the other group (Mean= 1.81). Participants of the first group might be more lenient with YDS washback effect because they are relatively under less stress than the others who have to prove their language proficiency in a specific time. Meanwhile, under no separate factor had gender and academic discipline a significant effect on participants' responses.

Considering the items related to foreign language test strategies, participants have mostly negative attitudes towards YDS since the test mostly motivates testees to learn test strategies and memorize some vocabulary items which are rarely used and less frequent words that are unlikely to be used in a real context, rather than speaking or using the foreign language communicatively. Academicians believe that YDS is a complex foreign language proficiency test that requires strategy training, mentor help and memorizing words that would never help testees learn and use the target language. Academicians who took this exam to have extra payment from the government or the ones who have professor or associate professor titles again responded to the related items more positively since they showed their foreign language competence in the past and do not want to criticize the test. Another interesting finding was the fact that most participants agreed on the merits of having a tutor to be successful in YDS. This single item could show the rationale of YDS itself. Learning some short cuts and basic question types could help test takers boost their performances in such tests, but could they speak or write well in that target language? Absolutely, they cannot because they spend their time with teachers practicing test techniques. Kitao and Kitao (1996) supported the fact that in the modern world, popular and reliable language proficiency tests should directly address test takers' communicative competence, the rest is all about the details on how much grammar or vocabulary knowledge they possess.

Another important issue in test washback effect was to check and see how motivating YDS could be on its test takers. Considering the items of AFTAS under the motivation factor, unlike what Külekçi (2016) stated in a part of his study in terms of the motivational effect of YDS, it was seen that participants disagree with the idea that YDS is a motivating test for academicians to learn a foreign language, and the test is highly difficult and does not contribute to participants' foreign language learning efforts at all. Heaton (1988) stated that if a test has considerably difficult questions for students to answer, the validity of that test (regardless of what it is made for) is debatable. If even native speakers of English have difficulty in taking high scores from YDS, what could be he the result for those who might be called "rookies" and strive for passing that language proficiency exam? They absolutely need to spend long hours to do it and take professional help from language teachers. Furthermore, test takers' perceptions about a test's negative or positive effects is another concern and to figure out to what extent people trust a particular test's positive effects could make that test a good test. Unfortunately, the items of AFTAS under the perception dimension showed that participants disagree with the idea that AFTAS has a positive effect on them and leads the academicians to learn and use the target language more effectively.

By responding multiple-choice questions, test takers do nothing but choose the best option which is impossible in real language use. They have to create their own utterances to conduct conversations when they really need that language. If the overall goal is to determine whether testees are able to use the target language effectively, how could it be possible to measure it with multiple-choice questions? Therefore, most academicians have difficulty in speaking and writing in a foreign language even if they passed YDS or its equivalent tests like KPDS or ÜDS (former language proficiency exams that were used in Turkey before 2013) in the past (Özmen, 2011). Finally, the last important effect of a test on test takers is its psychological effects. Determining if a test has positive or negative effects on people could give valuable data for decision makers since the psychological burden of an exam could demotivate learners and could possibly impede their efforts in language learning. When the test taker's psychology dimension was examined, it was found that the test causes stress and demotivates academicians in learning a foreign language. Yeşilyurt (2016) has also pointed out the same finding and reported that language proficiency tests like KPDS, ÜDS or YDS demotivates test takers and are mostly considered as obstacles to academicians' personal development and their future scientific efforts. Unfortunately, most academicians in Turkey are not lucky enough to spend some years abroad for language learning, that is why they mostly have limited backgrounds in language skills. Regarding this deficiency, they spend a lot of time studying grammar or memorizing vocabulary items, which would rarely help them in their academic lives.

6. Conclusions

In sum, language assessment is a multi-dimensional process and it should be planned cautiously since the outcomes might affect the whole society in the long term. Similar to the findings of the studies by Akın (2016) and Yavuzer and Göver (2012), the findings of this study revealed that YDS which aims to measure test-takers' foreign language proficiency levels has negative washback effect on test takers. Since this test is a classical multiple-choice exam and prepared to measure reading, vocabulary and grammar knowledge in a particular language and might in a way direct language learners to value memorizing specific grammar rules, academic vocabulary and practice complex reading skills. However, those test takers are also supposed to speak and write effectively in the target language, and those exams unfortunately misdirect the learners into mastering the mechanical skills of those languages. If language tests cause stress and weariness on language learners, it could be impossible to claim that these tests are good tests and serve for their educational functions. In this sense, this study clearly revealed that YDS's content and question types are not functional and should be reconsidered by the decision makers in Turkey.

In addition to the points mentioned above, the following suggestions that might assist future researchers on test washback issue. To begin with, the scope of this study was YDS, and there are some other international tests in Turkey to measure foreign language proficiency levels. Research concerning those tests' washback effect could also be carried out and check if those tests have positive effects on Turkish language learners. Next, most of the high-stakes tests in Turkey are multiple-choice tests, and future studies are necessary to find out the effectiveness of these tests and their washback effect on

Turkish students. Finally, a qualitative research to collect decision makers' and test writers' views on those tests' washback effect they observe could be conducted to obtain a deeper insight on this phenomenon.

References

Airasian, P. (1994). *Classroom Assessment* (pp. 18-21). New York: Mc Graw-Hill.

Akın, G. (2016). Evaluation of national foreign language test in Turkey. *Asian Journal of Educational Research*, *4*(3), 11-21. Retrieved in March, 2019 from: http://www.multidisciplinaryjournals.com/wp-content/uploads/2016/04/FULLPAPER

Alderson, J.C. (1986). Innovations in language testing. In M. Portal (Ed.), Innovations in language testing: *Proceedings of the IUS/NFER Conference* (pp. 93-105), Windsor.

Alderson, J.C., & Wall, D. (1986). Does washback exist? *Applied Linguistics*, *14*, 115-129.

Brown, J. D. (1998). *Testing in Language Programs*. Upper Saddle River. NJ: Prentice Hall Regents.

Büyüköztürk, Ş. (2013). *Sosyal bilimler için veri analizi el kitabı*. Ankara: Pegem.

Cheng, L., & Curtis, A. (2004). Washback in Language Testing. Retrieved in May, 2019 from: https://www.researchgate.net/publication/277405452_Washback_in_language_testing_Research_contexts_and_methods

Ching, P. Y. (2009). A review of washback and its pedagogical implications. *VNU Journal of Science Foreign Languages*, *25*, 257-263.

Cinkara, E., Tosun, Ö. (2017). Face validity study of a small-scale test in a tertiary level intensive EFL program. *B.U. Journal of Faculty of Education*, *6(*2), 395- 410.

Cohen, R. J., & Swerdlik, M. E. (2013). *Psikolojik test ve değerlendirme, testler ve ölçmeye giriş* (E. Tavşancıl, Trans.). Ankara: Nobel.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. USA: CBS Collage Publishers Company.

De Vellis, R. F. (2014). *Ölçek geliştirme, kuram ve uygulamalar* (T. Totan, Trans.). Ankara: Nobel.

Doğan, İ., & Doğan, N. (2014). *Adım adım çözümlü parametrik olmayan istatistiksel yöntemler*. Ankara: Detay.

Fish, J. (1988). Responses to mandated standardized testing. *Educational Researcher*, *18*, 27-32.

Green, A. (2007). IELTS washback in context: Preparation for academic writing in higher education. Retrieved in April, 2019 from: http://hosted.jalt.org/test/PDF/Dunkley2.pdf

Green, S. B., & Salkind, N. J. (2008). *Using SPSS for Windows and Macintosh: Analysing and Understanding Data*. Upper Saddle River: Pearson; Prentice Hall.

Hatipoğlu, Ç. (2016). The impact of the university entrance exam on EFL education in Turkey: Pre-service English language teachers' perspective. *Procedia-Social and Behavioural Sciences*, *232*, 136-144.

Heaton, J.B. (1988). *Writing English Language Tests*. Longman Handbook for Language Teachers. Longman.

Kitao, S. K., & Kitao, K. (1996). Testing communicative competence. *The Internet TESL Journal*, *2*(5), 1-6.

Kline, R. B. (2005). *Principles and practice of structural equation modelling*. New York: The Guilford Press.

Külekçi, E. (2016). A concise analysis of the Foreign Language Examination (YDS) in Turkey and its possible washback effects. *International Online Journal of Education and Teaching*, *3*(4), 303-315. Retrieved in March, 2019 from: http://iojet.org/index.php/IOJET/article/view/141/143

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, *29*, 4-16.

McEwen, N. (1995). Introducing accountability in education in Canada. *Canadian Journal of Education*, *20*, 1-17.

Noble, A.J., & Smith, M. L. (1994). *Measurement-driven reform: research on policy, practice, repercussion*. Tempe: Arizona University.

Özdamar, K. (2013). *Paket programlar ile istatistiksel veri analizi*. Eskişehir: 2013.

Özmen, K.S. (2011). Washback Effects of Inter University Foreign Language Examination on Foreign Language Competences of Candidate Academics. *Novitas Royal (Research on Youth and Language)*, *5*(2), 215-228.

Pearson, I. (1988). *Tests as levers for change.* Practice and evaluation. London: Springer. (pp. 98-107).

Polat, M. (2018). Akademisyenlerin Yabancı Dil Bilgisi Seviye Tespit Sınavı'nın (YDS) Geçerliğiyle İlgili Tutumlarının Belirlenmesi. Retrieved in March, 2019 from: https://dergipark.org.tr/download/article-file/693327

Şeker, H., & Gençdoğan, B. (2014). *Psikolojide ve eğitimde ölçme aracı geliştirme*. Ankara: Nobel.

Shohammy, E. (2006). *Language Policy. Hidden Agendas and New Approaches*. New York: Routledge.

Siegel, S. (1977). *Davranış bilimleri için parametrik olmayan istatistikler* (Y. Topsever, Trans.). Ankara: Ankara Üniversitesi Basımevi.

Smith, M. L. (1991). Put to the test: The effects of external testing on teachers. *Educational Researcher*, *20*, 8-11.

Vieira, A. L. (2011). *Preparation of the analysis. Interactive LISREL in practice*. London: Springer.

Wall, D. (1996). Introducing new tests into traditional systems: Insights from general education and from innovation theory. *Language testing*, *13*, 334-354.

Wall, D. (2000). The impact of high-stakes testing on teaching and learning: Can this be predicted or controlled? *System*, *28*, 499-509.

Yavuzer, H., & Göver, D. H. (2012). Akademik personelin yabancı dil durumu ve yabancı dil sınavlarına bakısı: Nevsehir örneği. *NEÜ Sosyal Bilimler Enstitüsü Dergisi*, *1*(2), 136-158.

Yesilyurt, S. (2016). An attempt for the exploration of academicians' experiences of the standard foreign language tests held in Turkey through metaphors. *International Journal of Higher Education*, *5*(2), 263-274.

Appendix A.

A.1. Academicians' foreign language test attitude scale (AFTAS)

Item	Strongly Disagree	Disagree	Partly agree	Agree	Strongly agree
1. YDS does not include listening questions that is why listening comprehension skills are disregarded in this test.					
2. YDS makes the test takers feel that learning a foreign language is easy.					
3. YDS includes many grammar questions that lead people study grammar more while learning a foreign language.					
4. Test takers are under stress before and after YDS.					
5. To be successful in YDS, a number of test techniques and strategies should be practiced.					
6. YDS motivates the test takers to learn a foreign language.					
7. YDS affects test takers negatively while preparing for the test.					
8. YDS affects people negatively in terms of language learning.					
9. Since YDS has advanced vocabulary questions, memorizing new words is a common way of language learning.					
10. YDS positively contributes to learning a foreign language.					
11. Test takers who study for YDS could use the target language effectively.					
12. YDS demotivates test takers.					
13. To have a tutor for YDS preparation can be a good idea.					

Türkiye'de uygulanan yabancı dil seviye tespit sınavı YDS'nin sınav etkisi: Bir ölçek geliştirme çalışması

Öz

Türkiye'de yabancı dil seviye tespiti için uygulanmakta olan YDS'nin (Yabancı Dilbilgisi Seviye Tespit Sınavı) adaylar üzerindeki olası sınav etkilerinin araştırılması bu sınava girenlerin sınavla ilgili ne hissettikleriyle alakalı değerli bilgiler verecektir. Bu çalışmada, YDS'nin olası sınav etkisini araştırmak amaçlanmış ve bunun için bir ölçek geliştirilmesine karar verilmiştir. Bu amaçla, ilk etapta içerisinde İngiliz dili eğitimcileri ve ölçme değerlendirme uzmanlarından oluşan 6 uzmanın katkılarıyla veri toplamada kullanılması planlanan ölçeğin taslak hali hazırlanmıştır. Hazırlanan bu taslak ölçek, bir sonraki safhada açıklayıcı ve doğrulayıcı faktör analizlerine tabi tutulmuş, geçerliği ve güvenirliği hesaplanmıştır. 3. ve son aşamada ise analizler sonrasında 13 maddeden ve 4 alt boyuttan oluşan, geçerliği ve güvenirliği sınanmış bu yeni ölçek marifetiyle Türkiye'deki 84 devlet üniversitesinden toplam 2683 akademisyenin YDS'nin sınav etkisi ile ilgili görüşleri toplanmış ve analiz edilmiştir. Sonuçlar, YDS'nin araştırmada kullanılan ölçekte belirlenen sınav stratejisi, motivasyon, algı ve adayların psikolojik durumları gibi alt boyutları dikkate alındığında, akademisyenlerin üzerinde negatif bir sınav etkisine sahip olduğunu ortaya koymuştur.

Anahtar sözcükler: sınav etkisi; yabancı dil ölçümü; merkezi sınavlar; YDS; ölçek geliştirme; açıklayıcı faktör analizi; doğrulayıcı faktör analizi

AUTHOR BIODATA

Murat Polat holds a Ph.D. at Osmangazi University, Institute of Educational Sciences, Research Methods and Statistics Program. Currently he is working as a language instructor at Anadolu University, School of Foreign Languages. His research interests include language assessment, alternative assessment methods, and educational statistics.