Journal of Educational Technology & Online Learning
Volume 4 | Issue 4 | 2021
http://dergipark.org.tr/jetol

Examination of a computer-delivered English proficiency test for flight attendants: Test-takers' experiences

Sezen Arslan [a] *

[a] Van Yüzüncü Yıl University, Turkey

Suggested citation: Arslan, S. (2021). Examination of a computer-delivered English proficiency test for flight attendants: Test-takers' experiences. *Journal of Educational Technology & Online Learning*, *4*(4), 673-687.

Article Info

Keywords:

Language testing
Civil aviation
Computer-delivered tests

Research Article

Abstract

International Civil Aviation Organization recognizes English proficiency tests for the aviation personnel. Despite this, no universal test has yet to be developed for this purpose. There are only local tests available on the market. Among them is a computer-delivered English proficiency test which is administered for a globally recognized flight company in Turkey. This test is implemented for flight attendants and has important consequences such as suspending flights until the required level of English is achieved in the test. Thus, test-takers' perceptions of the test are significant as they experience these impacts first-hand. This study, therefore, explored flight attendants' test-taking experiences. 26 flight attendants participated in this study. Semi-structured interviews and two focus group sessions were held to collect data and evidence. The findings showed that the respondents had concerns for test administration including the issues such as physical conditions of the setting, timing of the test, and test format. Although they agreed that the test had a representative sampling of the language, they thought it lacked relevant components in accordance with their needs and level. They also reported the test content did not reflect work-related context. The study, therefore, provided implications and suggestions to improve the test considering the aviation context.

1. Introduction

As highlighted in International Civil Aviation Organization (ICAO, 2009), English proficiency, especially the interaction skills of the aviation personnel are of great significance. This emphasis is understood from ICAO's regulations of English proficiency tests. Accordingly, speaking and listening skills should be involved in the tests rather than measuring grammar or vocabulary knowledge (ICAO, 2009). Thus, improvement of these skills is important for aviation personnel, particularly for flight attendants, since they are the first encounters for the passengers in the cabin. They are required to enable safety and security and communicate well with the passengers by listening to them actively, speaking clearly, and suggesting solutions to in-flight problems. Given that the passengers are often international, they should do these responsibilities while interacting in English. This highlights the need for them to have improved listening and speaking skills.

To explore whether/to what extent the flight attendants have this language performance required for the aviation settings, language testing is paramount. Thus, ICAO (2009) states that it recognizes English proficiency tests for the evaluation of language skills. It also maintains that these tests have high stakes as

* Corresponding author: Department of English Language Teaching, Van Yüzüncü Yıl University, Turkey
e-mail address: sezenarslan@gmail.com
This study was partly presented as a proceeding at the 1st International Conference on Educational Technology and Online Learning Conference held between 22-24 September 2021.

Doi: http://doi.org/10.31681/jetol.1003051
Received 30 Sep 2021; Revised 25 Dec 2021; Accepted 27 Dec 2021
ISSN: 2618-6586 . This is an open Access article under the CC BY license.

they have important consequences in terms of the safety and future of the industry. For this reason, the tests should adhere to the highest testing standards (Alderson, 2010). However, as emphasized by ICAO (2009), no standard universal test has yet to be developed; instead, regional or local testing procedures have been implemented for measuring English proficiency. As a result, it is of utmost importance to examine the quality of the language tests that are administered currently (Farris, 2016). Among them is a computer-delivered English proficiency test for flight attendants, which is implemented by a Turkish flight company with a globally eminent status. However, no studies have investigated this test in Turkey. This study fills this research gap by exploring the flight attendants' self-reported test-taking experiences concerning the test. In so doing, implications are provided throughout the study focusing on the validity, reliability, and authenticity of the test.

2. Literature

2.1. General Issues for Language Testing

Certain aspects are related to the quality of language tests. Among them are validity, reliability, and authenticity (Bachman & Palmer, 1996). According to Hughes (2003), a test is valid if it reflects a good sampling of the language skills, forms, or other aspects that it means to measure. Otherwise, the decisions based on the test results will not be accurate (Green, 2014). The other important aspect that highly impacts the quality of language tests is reliability. Bachman and Palmer (1996) explain reliability as the consistency of test results and maintain that potential factors that cause inconsistency should be minimized. Among those factors are physical conditions of the testing place, time of the testing, test-takers' background knowledge of the test content (Bachman & Palmer, 1996), test-takers' feeling fatigued, and poorly written test items (Neukrug & Fawcett, 2015). These factors may cause differences in the test performance across test-takers, thereby generating test bias (Bachman, 1990).

Another concern for language testing is the authenticity that is related to the "relationship between test task characteristics, and the characteristics of tasks in the real world" (Fulcher & Davidson, 2007, p.15). These tasks are specifically developed to mirror authentic activities that the test-takers might be required to come across in real life (Shohamy, Or, & May, 2017). For example, filling a form, writing a report, and simulations are deemed authentic test tasks. Bachman and Palmer (1996) relate this concept to the Target Language Use (TLU), thereby incorporating test-tasks that address a particular purpose and specific language use domain. For example, within the context of this study, several TLU tasks can be identified for flight attendants such as making announcements, writing briefings, telling emergency procedures, and offering meals. These tasks have a language-specific domain and work-related characteristics.

2.2 Issues for Language Testing in Aviation Context

Thanks to globalization, English has become the lingua franca of the world. Thus, it has become the intermediary language that enables communication among individuals from diverse language backgrounds. This eminent status of English coupled with the growth in the aviation industry and competition in the market to ensure safety and provide better service has made communication in English significant for aircraft personnel (Aiguo, 2007). Especially the flight attendants are required to use the language efficiently because they have various responsibilities in the aircraft: Taking care of handicapped and/or sick passengers, dealing with unexpected problems because of delays, in-flight meals, and other passenger complaints (Cornwall & Srilapung, 2013). In so doing, they are expected to communicate in English while providing service of high-quality for international passengers. This is of great concern for flight companies to have a leading role in the airline industry in such a competitive context.

Flight attendants should have a 'sufficient' English language proficiency to maintain a good interaction in the aircraft to offer the best service to the passengers. Therefore, there is a need for a language test to provide proof of their English proficiency. In response to this issue, ICAO introduces international language proficiency requirements and accordingly requires test-takers to demonstrate language skills in speaking and listening (ICAO, 2009). Although it is deemed a valuable endeavour for aviation language testing, there are some issues to be considered such as the quality of the tests (Farris, 2016) because there is no universal system of language testing in the aviation industry but there are many local language testing certification programs. Therefore, a question mark hovers over the validity and reliability of these tests implemented in different places in the world. For this reason, in this study, particular emphasis is given on providing implications for test developers to improve the validity, reliability, and authenticity of the English proficiency tests implemented for flight attendants.

3. Context of the Study: A Computer-Delivered English Proficiency Test for Flight Attendants in Turkey

ICAO only recognizes proficiency tests for measuring English language skills of the personnel in the aviation community and these tests aim to assess the test-takers' performance on listening and speaking abilities (ICAO, 2009). In response to this issue, one of the eminent Turkish flight companies developed an English proficiency test for flight attendants.

Flight attendants are classified into their ranks and responsibilities: The chief flight attendants (CFAs) are in the leader position in the cabin and are responsible for managing the cabin crew to enable safety and delivery of good service to the passengers (Dibakanaka & Hiranburana, 2012). Similarly, cabin attendants (CAs) are required to perform safety and security procedures and assist passengers. CFAs and CAs in a Turkish flight company are required to take a computer-delivered English proficiency text periodically to offer evidence of their English proficiency. The venue for this test is the test centre affiliated with the Turkish flight company in Istanbul, Turkey. A group of CFAs and CAs are accepted in this test centre and they are placed in test sessions where they take the test using a computer. Items in the test are randomized so that each test-taker is given a different set of test items. The perfect score for the test is 100 points. If the test-takers get 50 points and below, the company implements several sanctions such as suspending one's flights. When the test-takers get scores between 90-100 points, they are required to take the test after three years; however, for the points between 70-89, the test-takers are asked to sit for the exam after two years (Çelebi, n.d).

This test consists of a total of 11 questions in four sections and takes 25 minutes. The test-taker is allowed 15 seconds for thinking time and 75 seconds for giving a response to each question. The test mainly measures listening and speaking and it includes the following sections (BirebirİngilizceBirebirİngilizce, 2018; Çelebi, 2018):

(1) General section: There are five questions in this section. For each question, the test-takers are given 75 seconds to respond to the computer prompts. The questions are open-ended and they focus on general issues from real-life (e.g., *Which one is important for you- professional career or family life?, What are the advantages and disadvantages of studying abroad?).* The intended aim is to evaluate whether/to what extent the test-takers understand the prompts and respond to them orally by paying attention to pronunciation and appropriate vocabulary.

(2) Picture description: Two pictures are shown to the test-takers on the computer screen and they are asked to describe each picture in 75 seconds. Test-takers are asked to provide as many details as possible about the pictures. The pictures can be related to any context such as animals, people, and holidays. For instance, test-takers can be shown a picture where the people are celebrating an event and they are asked to list some details such as telling what they are doing, how they are feeling, and what kind of clothes they are wearing. In so doing, the intended aim is to evaluate the test-takers' speaking skills through considering pronunciation, fluency, grammar and task completion.

(3) Scenarios: The test-takers are provided with two scenarios where they are asked what they would do in the described situations. The test-takers are required to respond to each computer prompt in 75 seconds. Such a sample scenario can be given: "*You ordered food through a digital platform and when it was delivered, you realized that it went bad. What would you do*?" The intended aim of such items is to evaluate whether/to what extend the test-takers understand the prompts and respond to them. In so doing, the test-takers should provide a reasonable oral answer by paying attention to pronunciation and correct forms.

(4) Re-telling: Two reading passages are reflected on the computer screen. These passages have 90-120 words and they are about news or current affairs. The test-takers are asked to read the passages and re-tell them with their own words in 75 seconds.

Overall, the intended aim of the test is to evaluate whether/to what extent the test-takers recognize and understand the prompts by reading and listening and respond to them through demonstrating the speaking skills. Accordingly, the test-takers are evaluated on pronunciation, vocabulary, comprehension, fluency, structure, and interaction (Çelebi, 2018).

Considering that this test has serious professional consequences for flight attendants, this present study aims to investigate their test-taking experiences as they experience these impacts first-hand. In doing this, general principles for language testing (validity, reliability, and authenticity) are focused. This study is therefore deemed significant in that it also explores how this test can be improved in terms of these principles.

4. Methodology

4.1. Participants

A total of 26 (8 males, 18 females; age range: 23-45) flight attendants who work at a Turkish flight company were recruited for the study. Convenience and snowball sampling techniques were used respectively in the selection of the participants. First, the researcher found some members of the population from her network based on their availability; then, the snowball sampling technique was adopted by reaching out to further participants. Snowball sampling was assumed to be appropriate to this study as it facilitated obtaining rich information from different participants (Patton, 2014); thereby, overcoming problems concerning data collection from hard-to-reach populations (Etikan, Alkassim, & Abubakar, 2016; Sadler, Lee, Lim, & Fullerton, 2010).

The participants included CFAs (n=9) and CAs (n=17). They have a graduate degree in a wide range of fields including Civil Aviation, English Language Teaching, American Language and Culture, Chemistry, Accounting, Business, and Administration. Written consent was received from all participants for participating in this study.

4.2. Data Collection Tools

4.2.1. *Semi-structured interviews*

Semi-structured interviews were held with the participants to develop a deeper insight into the respondents' test-taking experiences (McGrath, Palmgren, & Liljedahl, 2019). The piloting of the interviews (n=10) resulted in minor changes in wording (See Appendix A). Interviews were conducted in the participants' native language (Turkish) for promoting a comfortable expression of their perceptions.

4.2.2 Focus groups

Apart from semi-structured interviews, two focus group interviews were also conducted. The main purpose in using focus group sessions was that they encouraged group interaction and provided opportunities of exploring why and how the respondents thought (Kitzinger, 1995). Focus group sessions were held in two

groups: The first group interviews were conducted with CFAs (*n*=5) while the second group interviews were conducted with CAs (*n*=12). Administering focus group interviews with two separate groups was because both groups were different in terms of years of experience and number of sitting for the language tests in the company; as a result, they might have different perspectives, thereby enabling discussion of various aspects of the test. Focus group interview sessions were held in Turkish after semi-structured interview analyses were done. They were moderated, recorded, and transcribed by two researchers (including the author).

4.3. Data Analysis

Thematic analysis was employed for investigating the qualitative data in this study. All interviews were transcribed verbatim and translated. The translated data were then checked by an independent researcher who is bilingual in Turkish and English to ensure a correct translation. The data analysis was made in two rounds. In the first round, semi-structured interviews were analyzed. Two researchers including the author worked on the analysis. In the first cycle, each researcher made an initial analysis of the data to generate data segments. Then, these segments were read and each researcher wrote analytic memos when they coded the data (Saldana, 2013). In the second cycle, these codes and memos were grouped and classified into certain themes. The researchers gathered to check the agreement/disagreement with this categorization of the codes into themes. In case of any disagreement on a code/theme, this disagreement was discussed till a final code/theme was produced. As a result, Cohen's kappa showed a substantial rate of agreement between coders (*k*= .802) (Landis & Koch, 1977). The final themes were used to investigate the data from the focus group sessions. Table 1 illustrates themes and subthemes with examples from flight attendants' words.

Table 1.

Themes, Subthemes with Illustrative Examples

Theme	Subtheme	Examples
Administration of the test	Testing conditions	Not enough silence; broken equipment
	Testing format	Human interviewer needed instead of computer prompts
	Timing of the test	Too early or late
Test content	Authenticity	The real use of English needed
	Range of structures and vocabulary	Wide range of structures and vocabulary covered
	The difficulty of test tasks	More visuals; questions with related background

5. Results

This section points out the key findings derived from the qualitative data in interviews and focus group sessions. Illustrative quotations from the flight attendants are also presented.

5.1. Administration of the Test

The qualitative data obtained from the interviews and focus group sessions were investigated to find out how the flight attendants felt about the administration of the computer-delivered English proficiency test. Accordingly, subthemes based on the data were: *a) Testing conditions*, *b) Testing format*, and c*) Timing of the test.* Results on each subtheme are presented below and pseudonyms are used throughout this study for confidentiality.

5.1.1. Testing conditions

All respondents (*n*=26) in this study reported that they had some issues related to the test administration. Among them, testing conditions were the most referenced aspect. The participants mentioned that there was not enough silence in the testing setting as each test-taker was required to answer the questions that were directed by the computer within the given time. They often indicated that there was a lot of noise when everyone spoke at the same time and that this was aggravated due to the headphones that were not noise-cancelling enough. They maintained that their performance was sometimes affected negatively as they were distracted by this noise and had difficulties in concentrating on the test tasks, as illustrated in the following excerpt:

"We are required to respond to computer prompts in a very short time. So, I try my best to collect my thoughts and pick the right words to utter. When I hear the test-taker speaking next to me, I really cannot focus on the questions and get anxious. Then, when the test ends, I realize that I could speak more about the tasks and feel sad for not showing my true performance." (Melis, CA)

For this reason, they noted that the headphones should have more effective noise cancellation, thus, the test-taker would not be disturbed by another test-taker sitting next to him/her while s/he was responding to the computer prompts. Additionally, they commented that as the test measured speaking skills and each test-taker was required to speak, external noise in the setting was understandable but it should be avoidable as much as possible. Therefore, they suggested that a small number of test-takers could be allowed to sit for the test in the same room and that they can be seated far apart from each other.

Apart from noise problems in the test setting, two participants also indicated in the focus group sessions that there could sometimes be broken equipment, which sometimes caused failure in the test. For example, one of the CFAs talked about her experience with broken equipment during the test. She said:

"At the beginning of the test, the computer system asks for testing the microphone whether it works properly or not. Although I tested the microphone and it worked properly, it was suddenly broken down towards the end of the test. Thus, I could not answer the questions and I failed the test." (Zeynep)

Along similar lines, another CFA also commented:

"Because of the sudden failure of the microphone, while I was once in the test, I could not show my true oral performance and got a lower score." (Ali)

As a result, testing conditions were reported to have a significant impact on the respondents' test performance. Among them, external noise and broken equipment were the most referenced factors.

5.1.2. Testing format

Many flight attendants (*n*=21) reported that they would prefer a human interviewer rather than a computer in the English proficiency test. In the focus group sessions, some respondents mentioned their reasons for disapproval of undertaking computer-delivered testing, as illustrated in the following excerpts:

"Responding to computer prompts is something very mechanic and dull. You just sit before a computer; it reads the questions for you and you are required to answer the question. Although the test measures speaking, it is not interactive in itself. Therefore, a human interviewer would enable more interactive assessment." (Kemal, CFA)

"I would prefer a face-to-face assessment because if I didn't provide enough answers to the questions, the human interviewer would change the topic or ask further questions. This would be a more realistic speaking assessment." (Orhan, CA)

"This computer test stresses me a lot. I always feel worried if things suddenly get out of control such as being disconnected from the system, equipment breakdown, etc. So, I think my speaking skills can be assessed by a human rater." (Dilara, CFA)

"The computer reads the question before I am expected to answer them. Then, a timer appears on the screen and the count-down starts. It is too mechanical and pressurizing. When I respond to the prompts, I always look at the timer. I get anxious. For this reason, a human interviewer would be more suitable as s/he won't be that pressurizing." (Deniz, CA)

"I become worried about any technical faults that can happen during the test. This fear distracts me a lot when I answer the questions in the test. If the test was administered by a human, I would be more relieved." (Nida, CA)

While a majority of the flight attendants (*n*=21) thought that they felt uncomfortable due to the computer-delivered tests, only a few respondents (*n*=5) indicated that they enjoyed responding to computer prompts during the test. They said they would not prefer human interviewers as they believed they would be pressurized. For instance, Gizem (CA) said:

"I would not like to be interviewed by a human. Speaking English gets more difficult when someone is looking at me, waiting for me to answer. This stresses me out. As a result, I cannot speak comfortably, which may have a negative impact on my oral skills. Now, I just take this test and it is more practical."

Echoing Gizem, Dogan (CFA) also says: "*I prefer computer-delivered tests as they are more practical and time-saving. It is also more convenient to read the questions on the computer screen and respond accordingly".*

The findings with illustrative prompts above mainly showed that most of the respondents preferred a human-to-human interaction during the test. They believed that a human interviewer would be more motivating in responding to the questions and make the testing procedures more authentic and interactive. Also, they noted that they feared potential technical problems that can happen during the test. Only some respondents indicated that they preferred computer-delivered tests as they believed these tests are more convenient and practical.

5.1.3. Timing of the test

Ten flight attendants in the interviews noted that the tests were sometimes administered at the very early or late time of the day. They thought that the timing of the test had a negative influence on their performance. For example, Can (CA) says:

"*We have a busy flying schedule, so we often need rest to refresh our minds and body. However, this test is sometimes administered at very early or late times of the day. This causes fatigue for me and leads my scores to decrease because I cannot concentrate easily."*

Similar to what Can expressed, Sibel (CA) also said:

"After long hours of flight, if the test is run very early next day, this stresses me more. I am afraid that I won't be able to get enough rest and answer the questions at those times. When I am extremely fatigued, I can't give my full attention to the test."

Interestingly, when this issue was brought into the agenda in focus-group interviews, most of the respondents (*n*=13) agreed that the time of the test was not an issue that directly impacted their test performance. For instance, Dilek (CA) expressed her opinion:

"*I think the timing of the test is not a problem for me because the test duration is short. Therefore, I can take such a short test regardless of its timing*."

In addition, Cemal (CFA) indicated a different viewpoint where he said: "*We are flight attendants. We are used to waking up early at night and late at noon as a requirement of our profession. Thus, it is not a big deal for me to take this test in the early or late time of the day*."

To conclude, among the flight attendants, there were two opposite viewpoints towards the timing of the test. Some of them reported that the early and late sitting of the tests had an impact on their test performance as they had already busy flight schedules and became increasingly fatigued. Therefore, they believed that preparing a test plan considering the flight schedules should be made to minimize fatigue. With a similar view, one of the CAs, introduced a solution concerning this issue during focus groups: "*If it is possible, we can take a day off before the testing day so that we can have rest."* (Mustafa). On the other hand, several respondents considered test timing suitable. As such, they reported that they were already used to a hectic life as a part of the profession and that the test did not take too much time to complete. Thus, they thought that they could sit for the test at any time.

5.3. Test Content

Concerning the test content, the flight attendants indicated some issues that can be grouped under the following subthemes: *1) authenticity, 2) range of structures and vocabulary*, and *3) difficulty of test tasks*.

5.3.1. Authenticity

The majority (*n*=22) reported that the test content did not reflect the real language use they mostly needed or experienced during the flight. During the focus-group sessions, some respondents commented that authenticity must be prioritized in the test, as illustrated in the following excerpts:

"Sometimes, English in the test does not reflect the real situations that we met during flight. For example, in the test, we can be asked whether/when someone tells lies as a general question, but this is not something directly related to our sector. Instead of it, there can be questions that are closely linked to the aircraft such as problems faced with the passengers." (Dilek, CA)

"As we are an international company, we carry many international passengers from every part of the world such as India, France, Germany, Africa…They all speak English with different accents. I sometimes find it difficult to understand their English. Therefore, some questions can include different accents, so we can develop a familiarity with them and we prepare for the test accordingly." (Ayhan, CFA)

"*We often experience issues with passengers such as delayed flights and meal service. Therefore, the test questions could include more of these cases. This will help us be prepared for these cases and when they occur, we can pick the most appropriate English words and expressions. In case of such cases during flights, we should address the problem as quickly as possible. So, there should not be any language problem."* (Ezgi, CA)

On the other hand, a couple of respondents (*n*=4) thought that the test content reflected an authentic use of language. One of them said:

"There are several items concerning the language use that we may face during the flight. Some items in the Scenario part of the test ask the test-taker what to do in a problematic situation in the cabin. These items are directly related to our profession." (Nilüfer, CA)

As a result, the majority of the respondents thought that the test tasks should reflect more potential problems they may face in the aircraft. They believed that these tasks could help them develop a familiarity with the appropriate words, structures, and expressions that can be used during those problems.

5.3.2. Range of structures/vocabulary

Range of structures and vocabulary was another most referenced subtheme that was reported as contributing to test validity. Accordingly, all of the respondents agreed that the range of English including grammar and vocabulary is quite comprehensive. One of the cabin supervisors during focus group sessions indicated that although he had taken this test several times, he felt he must study very hard before each test. He said:

"*In different parts of the test, there are questions with many structures and many groups of vocabulary like nouns, adjectives, adverbs. If you want to achieve in this test, you should master them. It measures speaking skills by including a wide range of language components (*Sadi, CFA).

Similarly, Beril (CA) echoed:

"*Even though this test measures oral skills, it touches on different parts of the English language as well. For example, there is a variety of grammatical structures from simple present to if-clauses. This is the same for vocabulary. You can find words regarding many topics. Therefore, you should know a good range of vocabulary to respond to the prompts in the test."*

Overall, the respondents felt that the test covered a large number of vocabulary and grammar items of English. Therefore, they believed that this was a positive side of the test as they reported that they needed to demonstrate a speaking performance by talking about a variety of topics with the passengers during the flight.

5.3.3. The difficulty of test tasks

The difficulty of the test tasks is another most cited subtheme in this present study. Responses are characterized by different aspects concerning the test task difficulty. Among them is the use of advanced vocabulary in the test (*n*=10). One of the respondents explained:

"*Especially the paragraphs can sometimes be difficult to understand because of advanced vocabulary. I don't think that preparing for high-level vocabulary contributes to my language level as I will not likely use them in my professional life*." (Selma, CFA).

Echoing this comment, another respondent mentioned: "*The vocabulary in the test items should be functional. If it is advanced and higher than my proficiency level, the test becomes intricate for me. Thus, it gets difficult for me to understand the items and provide a response*." (Sertap, CA)

In addition, several other respondents (*n*=11) agreed that some pictures in the picture-description section were not open for interpretation, which they thought might hinder talking further about the pictures. These thoughts were also captured in the focus group sessions, as illustrated in the following excerpts:

"*For example, think that there is a picture in the test. In the picture, there is a teacher with students in a room. You should look at the picture and talk about it, but I think there are not many details to talk about it except saying 'I can see a teacher and students. They are in a room'. But I am expected to talk much longer than this in the test. This makes me nervous*." (Ahu, CA)

"*The pictures should have been more open to interpretations. Some art pictures can be used to promote imagination, for example. There is always much to talk about them*." (Ada, CA)

Another aspect is linked to the use of field-specific words in the test content. Some respondents (*n*=6) explained that there were rarely technical words concerning other fields of study. For example, Cansu (CA) commented:

"In the test where I was once attended, there was a paragraph about diseases. I found it too difficult to comprehend the passage because I am not familiar with the medicine and field-specific words."

Overall, although the respondents indicated that the inclusion of a wide variety of vocabulary and structures in the test was beneficial for their oral skills because they had to study for a comprehensive list of them, they thought that the words used in the test should not be too advanced. In this aspect, the respondents considered the difficulty of test tasks in terms of the frequency and usefulness of the language based on their professional needs. They also suggested that more incentive pictures should be used to encourage them to speak further in English.

6. Discussion and Implications

This study investigated flight attendants' experiences of the computer-delivered English proficiency tests that they had to take periodically to provide evidence for their language proficiency. In doing this, the qualitative data obtained from interviews and focus group sessions were used. Accordingly, the main themes that emerged are 1*) Test administration (testing conditions, test format,* and *timing of the test*), and *2) Test content (authenticity, range of structures/vocabulary,* and *difficulty of test tasks*).

6.1. Administration of the Test

The findings in the present study showed that all respondents had some concerns about the administration of the test. Firstly, they noted that the testing setting had inappropriate conditions such as noise problems and broken testing equipment. They described instances where there was so much external noise that they hardly concentrated on the computer prompts and there were inconclusive tests due to the sudden breakdown of the microphone or headphones. As a result, they reported that these conditions had a negative impact on their test performance. This finding resonates with some studies that show noise is an influential distractor for test-takers (Kim, Baydar, & Greek, 2003; Mullis, Bohrnstedt, Preuschoff, Reyes, Stancavage, & Martin, 2012; Shu'Aibu,2021). Therefore, it is suggested that distracting sounds should be removed from the testing environment to ensure more reliable testing and assessment (Hughes, 2003). Equipment breakdown during the test is also a problem that impacts the test-takers' performance. According to Green (2014), it is one of the widespread problems occurring in testing practices. He considers this situation 'very paradoxical' because the test itself is not examined carefully while it aims to examine the test-takers' performance. Therefore, the testing equipment must be checked before the test is administered and necessary precautions must be taken. In case of any technical problem during the test, the invigilators must address the problem appropriately. Therefore, invigilators should also receive training for test administration and related problems.

Concerning test format, the majority of the flight attendants indicated their preference for human-to-human interaction instead of a computer-assisted language test. They believed that a human interviewer would make the test more interactive. This finding could be explained by the impact of personality traits on human-computer interaction (Pocius, 1991). Accordingly, introverted people can perform better in computer-related settings than extroverts as they do not need real people to interact with. Other respondents also mentioned that they feared technical faults that can happen during the test and they were distracted by this thought. This finding is in line with Schult and McIntosh's (2004) study which pointed out that the test-takers who were used to taking tests in traditional settings reported more anxiety about the thought of taking computer-assisted tests. Similarly, Taylor, Jamieson, Eignor, and Kirsch (1998) compared two groups of test-takers' English scores in a computer-assisted language test. The results showed that the group with more familiarity with computers had high test scores than the group with less familiarity. These similar findings could be explained by unfamiliarity with online testing or extra responsibilities undertaken during the test such as connecting to the system, running it properly, and taking the test online (Stowell & Bennett, 2010). Such findings highlight a need to develop the digital competence of the test-takers. Therefore, technological courses can be developed and provided for flight attendants for their in-service training. Also, they can be provided opportunities where they can periodically attend computer-delivered tests during their in-service training programs to develop a familiarity with this format.

Another remarkable finding concerning the test administration is the timing of the test. Whereas some respondents indicated that the timing of the test was not a big deal for them as they were already used to irregular work schedules as part of their profession, some of them mentioned that very early and late sitting of the tests was not appropriate for them as they felt they did not rest enough after long hours of flight

before taking the test. This finding shows the evidence of fatigue effect on the test-takers (Karimi & Biria, 2017) and calls for an appropriate timing in accordance with the flight schedules of the flight attendants.

To sum up, testing conditions including external noise, broken equipment, and timing of the test can impact the test reliability. As these conditions change from settings to settings, -that is not identical for every test-taker, there could be inconsistency in the test results. For example, in case of equipment breakdown, the test-taker's score is likely to be negatively impacted. Therefore, uniform conditions should be enabled by minimizing the differences between the sets of test administration (Hughes, 2003).

6.2. Test Content

Another main theme derived from the focus group sessions and interviews based on the flight attendants' experiences concerning the test was test content. It mainly includes aspects related to the authenticity of the test, the range of structures/vocabulary covered in the test, and the difficulty of test tasks. The findings revealed that most of the respondents thought the test tasks should reflect English that they can encounter in their professional contexts. This shows the significance of authentic tasks in testing languages (Hughes, 2003; Wigglesworth & Frost, 2017). Accordingly, the test tasks should represent authentic use of language that the test-takers may face in the professional life beyond the test. This finding points toward a need for TLU tasks (Bachman & Palmer, 1996) and the tasks that promote 'genuine interaction' (Lewkowicz, 2000). ICAO (2009) names these tasks as 'work-related' tasks and accordingly problem-solving activities, simulations, and briefings can be generated as tasks to test English proficiency for flight attendants.

English has become the intermediary language for people regardless of their language background (Erarslan, 2021). This points out the need to promote English as a lingua franca thanks to rapid globalization (Galloway & Numajiri, 2020). One of the respondents in this study highlighted this need as she agreed that it would be better to reflect other accents of English in the test as it can promote their awareness towards different accents. Given that the passengers have various accents of English, it is of crucial importance to consider English used in the aviation context as a lingua franca and to include the communication strategies for effective interaction with speakers from different language backgrounds (Kim & Elder, 2009). Therefore, an important implication for the test developers is that rather than focusing only on measuring English proficiency, communication strategy tasks should be integrated to evaluate whether/to what extent the test-taker deals with the problems in a work-related context. This could help prepare flight attendants for using English effectively in work-related problems.

Findings in this study also show that all of the respondents agreed that the test covers a wide variety of vocabulary and structures. The significance of wide sampling of English in terms of test validity is discussed (Green, 2014; Hughes, 2003). Accordingly, the test should represent the language as much as possible. Therefore, the test on which this study reports can be said to have content validation as it includes a proper sampling of the language. In so doing, the findings showed that it includes various functions of the language. If certain structures/words are only allowed in a test, the test may not be valid enough, thereby requiring very little test-taker preparation. This will end up with a negative backwash effect, which refers to the negative outcome of the test on how test-takers learn (Shohamy, Donitsa-Schmidt, & Ferman, 1996). For example, if the test recognizes a limited sampling of language components, the test-takers are likely to ignore the language components that are not covered in the test. Thus, it may have a negative impact on language learning.

Another remarkable finding of this study is that the majority of the respondents said some test tasks, especially, retelling tasks, were difficult for them. They thought that some passages were incomprehensible because of advanced words. This finding is not surprising as the respondents reported that there was a wide variety of vocabulary/structures in the test. Therefore, it is understandable that the level of vocabulary or structures might sometimes be more advanced than some test-takers' actual English proficiency. Therefore, this finding may imply the importance of the usefulness of test content according to test-takers' needs and experiences (Cobb & Laufer, 2021). According to Hughes (2003), a representative sampling of the language

is significant but the test should include the relevant components in accordance with the testing purposes. Therefore, he maintains that it cannot be expected for a test to include the same sampling for intermediate and advanced English learners. In this case, the test in the present study may include a proper sampling by recognizing the test-takers' language proficiency.

Concerning test difficulty, some other participants also indicated they sometimes had no/little prior knowledge about the passages; thus, they had difficulty in understanding and retelling the text. The positive contribution of prior knowledge to reading comprehension has been highlighted by several studies (e.g., Schuler, 2018; Yakut & Aydın, 2017). It is widely agreed that the prior knowledge of the topic helps the test-takers derive meaning easily and ease the comprehension. As highlighted by ICAO (2009), the language tests in the aviation industry should be designed to assess speaking and listening; it can thus be concluded that this test on which this study reports is not developed to assess reading. Therefore, reading passages can be added pictures to enhance comprehension (Canning-Wilson, 2001) or more comprehensible texts can be used for the flight-attendants so that they can be motivated to speak about them as the main aim of the test is to measure speaking skills. This can make the test more reliable as it minimizes advantaging certain groups of test-takers having prior knowledge of the text. Additionally, the test will become more valid as it focuses on encouraging test-takers to speak rather than relying on reading texts. Thus, it can measure what is intended.

Furthermore, for picture-descriptions tasks, a group of test-takers explained that some pictures used in the test were not open to interpretation; thus, they did not have too much to talk about them during the given time. This finding, therefore, sheds light on the design of the pictures in speaking skill tests for the aviation industry. Although pictures are non-verbal sources and are often used in the test to encourage the test-takers to speak, they should be designed carefully. In so doing, the pictures should be designed in a way to act as stimuli to elicit interpretative responses and infer and predict information (Canning-Wilson, 2001). It can also be suggested that the number of picture-description tasks can be increased to provide more chances of speaking for the test-takers, thereby improving the reliability of the test. As the number of observations increases in the test, a more precise understanding of the test-takers' performance can be obtained (Green, 2014).

7. Conclusion and Suggestions

Language testing in the aviation context has high stakes and is within an 'unregulated industry'; therefore, its adherence to good practices of testing by considering validity and reliability is of high importance (ICAO, 2009). For this reason, this study has contributed to the limited literature on language testing in this industry as it is the first study that examines flight attendants' first-hand experiences in Turkey. Drawing on the empirical research, the findings of the study have contributed to the expansion of the understanding of test takers' experiences considering reliability, validity, and authenticity in language testing. In so doing, the study shed light on the test design and administration. Thus, the present findings may be helpful for test developers for designing more authentic and field-specific English test tasks and mitigating the problems reported by the test-takers to improve testing conditions.

Although this study develops insights into the flight attendants' test-taking experiences, its limitations must be acknowledged. First, this study is a small-scale study conducted with a certain group of respondents; therefore, there can be a potential limitation with the transferability of the results. Future studies thus should focus on larger groups of respondents. Second, the study relied on the test-takers' perceptions, future studies thus should include the perceptions of test administrators and developers about the test design, implementation, and scoring. Third, this study employed convenience and snowball sampling techniques based on non-probability sampling. Although both techniques are useful in generating a sample and practical in collecting data, the sample may not be sufficiently representative of the target population thereby resulting in biased interpretation. Therefore, future studies with a similar focus are recommended to use purposive sampling.

References

Aiguo, W. (2007). Teaching aviation English in the Chinese context: Developing ESP theory in a non-English speaking country. *English for Specific Purposes, 26*, 121-128. https://doi.org/10.1016/j.esp.2005.09.003

Alderson, J. C. (2010). A survey of aviation English tests. *Language Testing, 27*(1), 51-72. https://doi.org/10.1177/0265532209347196

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

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

BirebirİngilizceBirebirİngilizce (2018). *DLA-Kabin ekibi dil sınavı [DLA-language test for flight attendants]*. https://www.birebiringilizce.com.tr/dla-kabin-ekibi-dil-sinavi.html

Canning-Wilson, C. (2001). Choosing EFL/ESL visual assessments: Image and picture selection on foreign and second language exams. *National Center for Research on Teacher Learning* (ERIC Document Reproduction Service No. ED452707).

Cobb, T., & Laufer, B. (2021). The nuclear word family list: A list of the most frequent family members, including base and affixed words. *Language Learning, 71*(3), 834-871. https://doi.org/10.1111/lang.12452

Cornwall, T. B., & Srilapung, V. (2013). Senior flight attendants' English communication needs: A case study of Thai Airways International. *US-China Foreign Language, 11*(4), 286-291.

Çelebi, D. (2018). *DLA teknikleri ve İngilizce bilgisi [DLA strategies and English].* Miran Yayıncılık.

Çelebi, D. (n.d). *DLA nedir? [What is DLA]*. https://thyingilizce.com/

Dibakanaka, A., & Hiranburana, K. (2012). Developing an e-learning competency-based English course module for chief flight attendants. *International Journal of Scientific and Research Publications, 2*(8), 313-326.

Erarslan, A. (2021). English language teaching and learning during Covid-19: A global perspective on the first year. *Journal of Educational Technology & Online Learning, 4*(2), 349-367. http://doi.org/10.31681/jetol.907757

Etikan I., Alkassim R., & Abubakar S. (2016). Comparison of snowball sampling and sequential sampling technique. *Biometrics & Biostatistics International Journal, 3(1),* 6-7. https://doi.org/10.15406/bbij.2016.03.00055

Farris, C. (2016). Aviation language testing. In D. Estival, C. Farris, & B. Molesworth (Eds.), *Aviation English: A lingua franca for pilots and air traffic controllers* (pp.75-91). Routledge.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book.* Routledge.

Galloway, N., & Numajiri, T. (2020). Global Englishes language teaching: Bottom-up curriculum implementation. *TESOL Quarterly, 54*(1), 118-145. https://doi.org/10.1002/tesq.547

Green, A. (2014). *Exploring language assessment and testing: Language in action*. Routledge.

Hughes, A. (2003). *Testing for language teachers.* Cambridge University Press.

International Civil Aviation Community (ICAO). (2009). *Language testing criteria for global harmonization*. https://www.ealts.com/documents/ICAO%20Cir%20318-AN180%20Global%20Harmonisation%20Testing%20Criteria%202008%20.pdf

Karimi, M., & Biria, R. (2017). Impact of risk taking strategies on male and female EFL learners' test performance: The case of multiple choice questions. *Theory and Practice in Language Studies*, *7*(10), 892-899. http://dx.doi.org/10.17507/tpls.0710.10

Kim, H., Baydar, N., & Greek, A. (2003). Testing conditions influence the race gap in cognition and achievement estimated by household survey data. *Applied Developmental Psychology, 23*, 567-582. https://doi.org/10.1016/S0193-3973(02)00142-9

Kim, H., & Elder, C. (2009). Understanding aviation English as a lingua franca perceptions of Korean aviation personnel. *ARAL, 32*(3), 1-17. https://doi.org/0.2104/aral0923

Kitzinger, J. (1995). Qualitative research: Introducing focus groups. *BMJ, 311*(7000), 299–302. https://doi.org/10.1136/bmj.311.7000.299

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. https://doi.org/10.2307/2529310

Lewkowicz, J. A. (2000). Authenticity in language testing: Some outstanding questions. *Language Testing*, *17*(1), 43-64.

McGrath, C., Palmgren, P. J., & Liljedahl, M. (2019). Twelve tips for conducting qualitative research interviews. *Medical Teacher, 41*(9), 1002-1006. https://doi.org/10.1080/0142159X.2018.1497149

Mullis, I. V. S., Bohrnstedt, G. W., Preuschoff, A. C., Reyes, I.D.L., Stancavage, F., & Martin, M. O. (2012). *Examining NAEP achievement in relation to school testing conditions in the 2010 assessments*. National Center for Education Statistics. https://www.air.org/sites/default/files/2021-06/NVS_Testing_Conditions_Paper_Final_0.pdf

Patton, Q. M. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Sage Publications.

Pocius, K. E. (1991). Personality factors in human-computer interaction: A review of the literature. *Computers in Human Behavior, 7*, 103-135.

Sadler, G. R., Lee, H.C., Lim, R.S.H., Fullerton, J. (2010). Recruitment of hard-to-reach population subgroups via adaptations of the snowball sampling strategy. *Nursing & Health Sciences, 12*, 369-374.

Saldana, J. (2013). *The coding manual for qualitative researchers*. SAGE Publications.

Schult, C. A., & McInthosh, J. L. (2004). Employing computer-administered exams in general psychology: Student anxiety and expectations. *Teaching of Psychology, 31*(3), 209-211. https://doi.org/10.1207/s15328023top3103_7

Schuler, J. (2018). Looking at and beyond the lexical surface in L2 reading comprehension: Insights from a video-based study. *Language Awareness, 27*(1-2), 113-135. https://doi.org/10.1080/09658416.2018.1435672

Shohamy, E., Or, L. G., & May, S. (2017). *Language testing and assessment*. Springer.

Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Test impact revisited: Washback effect over time. *Language Testing, 13*(3), 298-317. https://doi.org/10.1177/026553229601300305

Shu' Aibu, M. G. (2021). The effect of testing conditions on the students' performance of distance learning system (DLS) in Jigawa State. *International Journal of Contemporary Education Research, 20*(8), 71-80.

Stowell, J. R., & Bennett, D. (2010). Effects of online testing on student exam performance and test anxiety. *Educational Computing Research*, *42*(2), 161-171. https://doi.org/10.2190/EC.42.2.b

Taylor, C., Jamieson, J., Eignor, D., & Kirsch, I. (1998). *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. Educational Testing Service.

Wigglesworth, G., & Frost, K. (2017). Task and performance-based assessment. In E. Shohamy, L. G. Or, & S. May (Eds.), *Language testing and assessment* (pp.121-133). Springer.

Yakut, A. D., & Aydın, S. (2017). An experimental study on the effects of the use of blogs on EFL reading comprehension. *Innovation in Language Learning and Teaching, 1*(11), https://doi.org/10.1080/17501229.2015.1006634

Appendix A: Semi-structured interview protocol

1. Could you please provide information about your age, rank (cabin attendant or chief flight attendant), years of professional experience and department of graduation from the university?

2.What do you think about the design of the test?

2a. What do you think of the difficulty of the test?

2b. Which part of the test is most difficult for you? Why?

3. What do you think about the administration of the test?

3a. How do you feel taking this computer-delivered English proficiency test?

4. What do you think are the factors that have an impact on your test performance?

5. Do you have any further comments and/or suggestions?