Applied Linguistics Papers 27/1: 2023, 79–90

The Performance of Brazilian Air Traffic Controllers in Radiotelephony Communications in English: the Controllers' Perspective

Natalia DE ANDRADE
Department of Airspace Control
E-mail: nandraderay@gmail.com,

Abstract: Following the argument-based validity framework (M.E. Kane 2006, 2013), this paper aims at analyzing controllers' perspectives of their performance in radiotelephony communications in English, utilizing data generated through the ongoing validation process of EPLIS, the Aviation English Proficiency Exam for Brazilian Air Traffic Controllers. This paper reports one phase of a broader multistage mixed-methods study and focuses on the consequences inference. Part of the data generated from a 65-item questionnaire answered by Brazilian Air Traffic Controllers who got proficiency Level 3 (considered non-operational) and Level 4 (considered by ICAO as the minimum level necessary to control aircraft using the English language) is analyzed. The questionnaire was divided into 3 parts: test-takers' perceptions of the rating scale established by ICAO; test-takers' perceptions of their proficiency while controlling aircraft in English; and problems they faced while controlling in English.

Keywords: validity, assessment, Aviation English, radiotelephony communications, air traffic controllers' language

Introduction

English proficiency in the aviation context is one of the aeronautical safety measures, i.e. a protective layer to mitigate risks of incidents or accidents. In response to concerns that insufficient English proficiency on the part of non-native English speaker (NNES) pilots or air traffic controllers was a contributing factor in the chain of events leading to accidents or incidents, the International Civil Aviation Organization, henceforth ICAO, decided to strengthen provisions related to English for radiotelephony communication. In 2004, they published the first edition of the *Manual of Implementation of the ICAO Language Proficiency Requirements*, also called Doc 9835. The purpose of the document was to show Member States the importance of providing their staff with adequate proficiency "in conducting and understanding radiotelephony communications in the English language" (ICAO 2010: vii) and to establish measures related to training and testing that should be applied by the member countries.

As a Member State, Brazil has sought to comply with all ICAO recommendations by establishing an action plan to implement the Language Proficiency Requirements (LPRs). The Institute of Air Space Control (ICEA), an Institution of the Brazilian Air Force, was therefore designated by the Department of Air Space Control (DECEA) to develop the proficiency assessment to measure the proficiency of Brazilian air traffic controllers, named EPLIS.

Studies analyzing the criteria established by ICAO have been carried out in some parts of the world to better understand whether it is aligned with the reality of aviation communication. However, there is a lack of research projects aimed at analyzing the tests made available after the publication of ICAO`s policy, according to J.C. Alderson (2008: 6) and D. Estival et al. (2016).

Considering the high stakes of the aviation context, this paper presents the data and discussions on part of the second phase of a broader multiphase mixed-methods study and analyzes the performance in English of Brazilian air traffic controllers considering their perspectives. The aim of this paper is aligned with the importance of a careful analysis of the exams used to measure the proficiency of aviation professionals and the negative consequences of the use of instruments that are not suitable for their purpose, as indicated by J.C. Alderson (2008: 15).

1. ICAO language policy

Several language assessment theorists have analyzed and critiqued the ICAO LPRs. One of the main findings of the studies in relation to the language policy established by the ICAO is that proficiency in the international aeronautical context is limited to language knowledge (H. Kim/ C. Elder 2009). For instance, U. Knoch (2014) sought to elicit, through focus group interviews with pilots, indigenous criteria for the aeronautical context. The pilots who participated in the focus groups heard the performance of non-native English-speaking pilots and assessed them according to the operational needs. Pilots demonstrated difficulty in differentiating linguistic knowledge from operational knowledge, which reinforces the view that the construct established by the ICAO language policy is limited, as criticized by T. McNamara (2012). U. Knoch (2014) used the results of her analysis to question the validity of the ICAO Proficiency Scale and concluded that it presents construct irrelevant variance of the communication between pilot and air traffic controller. According to U. Knoch (2014: 108) the scale is not fully representative of what happens in radio communication.

Some countries have subverted the ICAO's recommendations. South Korea, for example, was one of the first countries to develop its own test of aviation English. However, it subverted the ICAO policy by disclosing all forms of the test on its website, thereby "ensuring that all aviation personnel were able to reach the minimum required standard" (H. Kim 2013: 103). In Korea, pilots and air traffic controllers had been licensed as level 4 and continued working with international traffic.

Other countries, as shown by T. McNamara et al. (2019: 195) decided to choose a simpler exam that all operators could pass or allowed multiple attempts by the examinees in the same exam. T. McNamara et al. (2019: 195) also point out that decisions like these, although shocking to some researchers in the field of assessment, are justifiable in this context, as they are fairer assessment practices, since older and more experienced operators would not achieve the minimum proficiency determined by ICAO, which would culminate in the loss of professionals with vast operational knowledge, who never had problems related to the use of the English language. T. McNamara et al. (2019) do not cite the case of Brazil, we can say that there is an ongoing national language policy that differs in part from the policy established by the ICAO. The following section explains how the ICAO language policy was established and is carried out in Brazil.

2. The national policy: an overview of EPLIS intended interpretations, uses, and consequences.

According to the *Aeronautical English Proficiency Requirements Manual* (2018), henceforth referred to as MCA 37-225, EPLIS is an aviation proficiency test developed to comply with ICAO`s requirements. The MCA 37-225 states that EPLIS scores predict the performance of Brazilian air traffic controllers while dealing with international traffic and therefore help ensure greater efficiency in radiotelephony communications decreasing the risk of accidents and incidents (ibid: 35).

The main use of the test is to license controllers who demonstrate the minimum level of proficiency required to work with international traffic. Nevertheless, as Brazil has a large territory and the largest air space in South America, the need for hiring air traffic controllers subsumes the possibility of training or hiring professionals whose proficiency in aviation English meets the requirements. For this reason, air traffic controllers are licensed even though they do not achieve the required aviation English proficiency level. Air traffic controllers who achieve level 3 or below are licensed to control both domestic flights and international flights when supervised by an experienced colleague whose proficiency is level 4 or higher.

MCA 37-225 establishes that once the Brazilian Air Force has enough operational level 4 or higher controllers, English proficiency will become a restrictive criterion to issue air traffic control licenses for international traffic control purposes. The regulation also lists the mitigation measures taken in the operational management to ensure safety in the international air traffic control service. These actions include a) distributing the controllers who were assessed as level 4 or above in EPLIS equally each shift and b) auditing shifts to analyze if the number of ATCO is high enough to ensure safety in the communication with international pilots.

In addition, MCA 37-225 also presents the aviation English training program available for Brazilian air traffic controllers. In accordance with ICAO`s policy, member states should provide ongoing training to both the ATCO who scored the minimum level 4 and to those who were not able to achieve the recommended score. In this case, the score in EPLIS is used both to determine which training the ATCO should receive and to better place the controllers in the aviation English courses provided. Table 1 below illustrates the uses and consequences of EPLIS scores, according to the MCA 37-225:

Uses	Consequences
U-1 Personnel Licensing	C-1 Ensuring safety in the Brazilian Air Space
U-2 Placement for air traffic shifts	C-2 Mitigation measures
U-3 Placement for aviation English courses	C-3 More focused training programs

Table 1. Uses and their intended consequences of EPLIS.

The main use of EPLIS's scores is personnel licensing as a means of ensuring safety in the Brazilian Air Space. For this reason, it is essential to analyze whether scores in EPLIS

account for the necessary communicative language ability air traffic controllers should demonstrate when controlling international traffic in English.

3. An argument approach to validation

M. Kane (2013) indicates that the process outlined within the argument-based approach encompasses the claims that are made in a proposed interpretation or use, also called the Interpretative Use Argument (IUA), and the evaluation of these claims, the Validity Argument. An interpretive argument references the proposed interpretations and uses of test results by systematizing the network of inferences and assumptions from the observed performance to the conclusions and decisions based on the test takers' performances. According to M. Kane (2006: 23), the interpretative argument involves inferences leading from observed performances to the claims based on these performances and each of these inferences should involve an extension of the interpretation or a decision.

As M. Kane's interpretive argument enables the reasoning inherent in the proposed interpretations and uses, it can be a solid starting point for both test developers and test evaluators. However, the author indicates that the IUA cannot be understood as a checklist, since it does not follow an established pattern and varies according to each test. M. Kane (2013: 12) develops S.E. Toulmin's (2003) approach to practical reasoning with claims, data, warrants, backing and rebuttal as the basis for an interpretive argument. In S.E. Toulmin's structure, the claim indicates the conclusions drawn about test takers based on observations of test takers' performance (data). The data consists of empirical observations (e.g. test scores) on which the argument is built. A warrant links the data to a claim, legitimizing the claim by showing the data to be relevant. Backing for an argument gives additional support to the warrant, whereas a rebuttal suggests a counter-argument to the claim.

M. Kane (2013: 13) indicates that the IUA should not be understood as an end in itself. It should be developed to make the evaluation of the proposed interpretation and use as rigorous as possible. He indicates that the IUA should be stated in enough detail to guide an effective validation. C.A. Chapelle, M.E. Enright, and J. Jamieson (2008) and U. Knoch/ S. Macqueen (2019) state that an argument approach to validation requires test developers and researchers to specify what is entailed in test interpretation and use. It is a means of justifying the meaning of scores and the inferences that are drawn based on test scores. The argument-based approach relies on specifying inferences, warrants and assumptions associated with score interpretations and uses.

4. Research methodology

This study is part of a broader multiphase mixed-methods research (J.W. Creswell/ V.L. Plano Clark 2011: 2018) and illustrates part of the quantitative phase of an argument-based validation study of EPLIS. Based on the uses and intended consequences of EPLIS presented in Table 1, this article aims to answer the following research question: According to Brazilian Air Traffic Controllers' perceptions, are the levels in EPLIS adequate to cope with non-routine situations in English in radiotelephony communications?

To answer the above research question, part of the quantitative analysis of a 65-item questionnaire is presented. The questionnaire was sent by email to all controllers who obtained level 3 in EPLIS 2018 or had valid levels 4, 5 and 6 in June 2019, totaling 1887

air traffic controllers. After one month, data were extracted from the Qualtrics program. Altogether, 704 respondents completed the questionnaire, which means a response rate of approximately 37%. However, only 701 respondents were considered for this research, as they fully completed the questionnaire.

The return rates of all proficiency levels were quite representative in relation to the number of controllers at each operational level. Therefore, the sample, despite having a statistically non-normal distribution, represents the real population. Among those who achieved proficiency level 3 in EPLIS, 176 responded to the questionnaire, which represents a response rate of 40%. The response rates varied from 32% among those who had a valid level 4, 37% among level 5 controllers and 42% among the professionals assessed as level 6. To analyze the data generated by the questionnaire, descriptive and inferential analysis in SPSS (non-parametric tests: W.H. Kruskal-Wallis/ W.A. Mann-Whitney) have been carried out.

5. Analyzing controllers' perceptions

In the questionnaire, there were four questions that elicited respondents' perceptions on their level of confidence when controlling foreign aircraft. Table 2 below summarizes the analyzed groups and illustrates, the cases in which there was a statistically significant difference in the controllers' confidence level in relation to their level in EPLIS and their experience as controllers.

Level of confidence	Across proficiency levels	Between L3 and L4	Across experience groups among L3	Across experience groups among L4
	p-value	*p*-value	*p*-value	*p*-value
Q33 Do you feel confident to interact with pilots using phraseology in English?	.000	.000	.895	.113
Q34 Do you feel confident to interact with pilots using plain in English when phraseology is not enough?	.001	.000	.508	.225
Q35 Do you feel confident when you interact with pilots who are native speakers of English?	.001	.000	.841	.095
Q36 Do you feel confident when you interact with pilots who are not native speakers of English?	.001	.000	.656	.275

Table 2. Controllers' perceptions about their level of confidence across different groups.

The responses on the R. Likert scale for the 4 questions could vary between: 1-not confident, 2-little confident, 3-confident, 4-very confident. Kruskal-Wallis tests were applied for each of the questions in Table 2, and the results indicated statistically significant

differences in controllers' confidence level across proficiency levels[1]. Respondents with a proficiency level of 4 or higher had an average response close to "very confident" (Md ≅ 4). In order to verify which groups had different response patterns, post-hoc W. A. Mann-Whitney tests were performed, comparing pairs. The results of these tests demonstrated that there was no significant difference between the groups with an operational level or higher, so it can be stated that the level of confidence declared when controlling in English is pretty much the same across the groups. In contrast, the average response among those who achieved proficiency level 3 in EPLIS was "confident" (Md ≅ 3). There was a significant difference in the W.A. Mann-Whitney test between levels 3 and 4 and, consequently, between the higher levels, as illustrated in the second column of Table 2 above.

Table 3 below summarizes the mean, standard deviation, and standard error values of controller responses at each proficiency level and exemplifies the response patterns for Questions 33 and 34. When asked about the level of confidence when controlling using only phraseology in English, the average response at all levels is slightly higher than when using plain English. By verifying the response patterns between levels 3 and 4 in these questions, using the W.A. Mann-Whitney test, it was possible to confirm statistically significant differences between the two groups (p-value = 0.000 in Q33 and p-value = 0.001 in Q34). These results reinforce that the higher the level in EPLIS, the more confident the controllers feel in radio communications in English, both in phraseology and in plain English.

	Subgroup	N	M	SD	SE
Q33 Do you feel confident to interact with pilots using phraseology in English?	L3	176	3.25	0.064	0.845
	L4	380	3.56	0.034	0.661
	L5	96	3.74	0.058	0.567
	L6	49	3.92	0.049	0.344
Q34 Do you feel confident to interact with pilots using plain in English when phraseology is not enough?	L3	176	3.01	0.064	0.855
	L4	380	3.36	0.038	0.740
	L5	96	3.61	0.065	0.639
	L6	49	3.88	0.056	0.389

Table 3. Controllers' perceptions about their level of confidence.

We observe, then, that all groups feel confident when controlling using the English language, but the degree increases from "confident" to "quite confident" in relation to the non-operational proficiency level (L3) and operational or higher (L4, 5 and 6). To ensure

[1] It has been considered a p-value less than 0.05 as significant, according to J. Pallant's (2013) definition.

the reliability of the data, the effects of sample size were calculated, considering the difference between the number of respondents at different levels of proficiency. In all W.A. Mann-Whitney tests performed, the effect of sample size was low ($r < 0.2$).

Experience, however, did not appear in the data as an impact on controllers' confidence when operating in English. The W.H. Kruskal-Wallis tests between experience groups at level 3 did not demonstrate a statistically significant difference in the confidence level, with the p-value above 0.5, as illustrated in the third and fourth columns of Table 3. In other words, it statistically demonstrates that highly experienced controllers who obtained level 3 in EPLIS declared the same level of confidence when controlling in English as the ones with less experience in the profession. The same occurred in the W.H. Kruskal-Wallis tests between experience groups at levels 4, 5 and 6.

In self-assessment section of the questionnaire, the controllers answered questions about their performance in aeronautical English in real life situations. The questions focused on the relationship between performance in EPLIS and performance in real situations of use, as well as performance on the ICAO established criteria. When asked if the level of proficiency in English would be sufficient to communicate with international pilots, respondents showed similar response patterns, as can be seen in Table 4 below. Answers could vary between: 1- not enough, 2- little enough, 3- enough, 4- quite enough.

	Subgroup	N	M	SD	SE
Q40 Is your proficiency level sufficient to communicate with international pilots?	L3	176	3.22	0.056	0.740
	L4	380	3.46	0.030	0.591
	L5	96	3.64	0.058	0.564
	L6	49	3.88	0.047	0.331

Table 4. Sufficiency of the levels in EPLIS in real life situations.

When comparing means and SD among proficiency groups, it was possible to verify that all groups believe that their proficiency level is sufficient to communicate in real situations in English, including controllers who were assessed as level 3 in EPLIS. The difference, however, lies in the variation between “sufficient” and “quite sufficient”. As illustrated in Table 4, response patterns varied across proficiency levels. Both the W.H. Kruskal-Wallis and the W.A. Mann-Whitney tests showed that there is a significant difference between the answers, since the higher the level of proficiency in EPLIS, the closer to “fairly sufficient” the average responses. Experience, once again, did not appear as something that had an impact on the controllers' response pattern. Mann-Whitney tests considering experience indicated no statistically significant difference in response patterns.

There were 9 questions about the performance of controllers in specific situations of use of the English language in radio communications. Responses could vary between 1 (never), 2 (rarely), 3 (often) and 4 (always). The response patterns across proficiency levels were quite similar, as can be seen in Table 5 below:

	Subgroup	N	M	SD	SE
Q43 Do pilots understand your instructions in English?	L3	176	3.66	0.037	0.487
	L4	380	3.70	0.024	0.460
	L5	96	3.77	0.046	0.447
	L6	49	3.90	0.044	0.306
Q44 Do pilots ask you to repeat your instructions when you are controlling in English?	L3	176	1.77	0.052	0.691
	L4	380	1.75	0.039	0.760
	L5	96	1.64	0.074	0.727
	L6	49	1.53	0.121	0.844
Q45 Do pilots readback correctly when you control in English?	L3	176	3.60	0.040	0.525
	L4	380	3.59	0.029	0.563
	L5	96	3.67	0.051	0.496
	L6	49	3.80	0.058	0.407
Q46 Do native English-speaking pilots understand your instructions in English?	L3	176	3.56	0.046	0.611
	L4	380	3.65	0.026	0.516
	L5	96	3.69	0.056	0.549
	L6	49	3.88	0.047	0.331
Q47 Do non-native English-speaking pilots understand your instructions in English?	L3	176	3.62	0.038	0.510
	L4	380	3.62	0.026	0.512
	L5	96	3.65	0.059	0.580
	L6	49	3.86	0.051	0.354
Q54 Do you think your pauses interfere with pilot's understanding when you communicate in English?	L3	176	1.95	0.055	0.727
	L4	380	1.99	0.044	0.852
	L5	96	1.85	0.088	0.858
	L6	49	1.65	0.147	1.032
Q59 Do you believe your interactions in English are successful?	L3	176	3.48	0.042	0.555
	L4	380	3.59	0.027	0.524
	L5	96	3.75	0.044	0.435
	L6	49	3.90	0.044	0.306

Table 5. Controllers' perceptions on their performance.

Questions 43 and 44 are complementary. In Q43, the average responses of level 3 controllers (Md = 3.66) range from "often" to "always" in relation to the pilots' under-

standing of their instructions when controlling in English. The low DV indicates the homogeneity of responses in the sample. Consequently, the average response of these controllers in Q44 was low (Md = 1.77), thus indicating that the frequency with which pilots ask them to repeat an instruction in English ranges from "never" to "rarely". In Question 43, the p-value in the W.H. Kruskal-Wallis test by level of proficiency was .005, which indicates a statistically significant difference between responses. However, when performing the W.A. Mann-Whitney test two by two between levels 3 and 4, the results indicated an insignificant difference between the responses, with p-value of .418 and sample effect size test at r = 0.0.

A statistically significant difference between proficiency groups was found when we compared levels 3 and 4 with levels 5 and 6 in the W.A. Mann-Whitney tests. The same trend occurred in the tests performed with Q44. In both Questions, the W.H. Kruskal-Wallis tests by level of experience at level 3 indicated statistically significant differences in the pattern of responses, with p-value established at .015 in Q43 and .010 in Q44. The average of responses by experience group revealed that the most experienced group of level 3 controllers considers that pilots always understand their instructions (Md = 3.98) and never ask to repeat an instruction in English (Md = 1.16).

Regarding Q45, the W.H. Kruskal-Wallis test plotted that there was no statistically significant difference between the response patterns by proficiency level (p-value = 0.06). In other words, regardless of the level of proficiency, the controllers stated that the pilots always/frequently make the correct readback in interactions in English; that is, they demonstrate, through this procedure, that they understood all the details of the instruction. The mean number of responses in Q45 was higher at level 3 (Md = 3.60) than at level 4 (Md = 3.59), but the difference is not statistically significant, according to the two-by-two W.A. Mann-Whitney test (p-value = 0.969) and the sample size effect test (r = 0.0).

Q46 and Q47 elicited whether controllers noticed any difference when controlling foreign aircraft. However, the difference between response means was not representative and all levels of proficiency indicated that the frequency in which native and non-native pilots understand instructions in English ranges from "often" to "always". The W.A. Mann-Whitney post-hoc tests between levels 3 and 4 in both questions showed no statistically significant difference in the responses of the two groups (p-value = 0.196 in Q46 and p-value = 0.961 in Q47), and the test of sample effect size was low (r=0.1 in Q46 and r=0.0 in Q47). Therefore, the data show that according to both the pre-operational level (L3) and the operational and higher levels (L4, 5 and 6) are understood by both native and non-native speaker pilots.

Question 54 elucidates the controllers' perception of the influence of their pauses in communication and the answers varied between "never" and "rarely" for all levels of proficiency, as illustrated in Table 5 above. When we compare the average responses of levels 3 and 4, we can see that the average of controllers with level 3 (Md = 1.95) is lower than that of level 4 (Md = 1.99) and the low DV demonstrates the uniformity of the data. However, the W.A. Mann-Whitney test demonstrated that this difference is not statistically significant (p-value = 0.918). There was also no statistically significant difference in the results of the W.A. Mann-Whitney test performed between level 3 and level 5 (p-value = 0.183). It is only when we compare the responses at levels 3, 4 and 5 with those of controllers with level 6 in EPLIS that the results of the W.A. Mann-Whitney tests reveal a

statistically significant difference; however, the average response at this level also varies between “never” and "rarely".

When asked whether their interactions in English are successful, the average of answers varied between “always” and “frequently”, with the lowest average of 3.48 for level 3 and the highest average of 3.90 for level 6. The W.H. Kruskal-Wallis test between proficiency levels indicated that there is a statistically significant difference in the pattern of responses (p-value = 0.000), so post-hoc W.A. Mann-Whitney tests were performed. The test between levels 3 and 4 showed a statistically significant difference, as the p-value was 0.037. The sample effect size test was low (r = 0.1). However, even though they were statistically different, the response pattern was positive for both levels. The W.H. Kruskal-Wallis test between the experience groups at level 4 showed that there is a statistically significant difference in the response pattern (p-value = 0.037) and, the more experienced, the greater the success in the interaction reported by the controllers. It can be concluded, then, that all levels of proficiency consider their interactions in English with international pilots successful.

When asked if they had experienced an air traffic control situation in English that they could not solve by themselves (Q62), most respondents, regardless of their proficiency level, indicated “no” (82.19%). However, among those who answered “yes” (17.81%), proficiency ranged from level 3 to level 6 in EPLIS, with the highest percentage among L3 (23.9%) and the lowest among L6 (4.8%). However, even with low percentages, it is important to note that even between levels 5 (10.4%) and 6 (4.8%), there are control situations in English in which respondents indicate that they needed help to resolve.

	Subgroup	N	No	Yes
Q62 Have you experienced a situation which you could not solve by yourself?	L3	176	76,1%	23,9%
	L4	380	86,1%	13,9%
	L5	96	89,6%	10,4%
	L6	49	95,2%	4,8%
Q64 Have you ever experienced a situation in which phraseology was not enough?	L3	176	58,0%	42,0%
	L4	380	47,9%	52,1%
	L5	96	38,5%	61,5%
	L6	49	48,2%	51,8%

Table 6. Real life situations experience.

In Q64, the percentage of respondents who indicated that they had already experienced situations in which the standard phraseology was not enough was 61.21%, while 38.79% stated that they had not experienced this type of situation. In this question, the percentages of those who answered “yes” among controllers with level 4 and level 6 in EPLIS were quite similar, being 52.1% among L4 and 51.8% among L6. Among controllers with level

3 on EPLIS, 42% said they had already experienced a situation where standard phraseology was not enough, which was the lowest percentage among the different proficiency levels.

Conclusion

The analysis of air traffic controllers` perceptions of their performance has shown that there is statistically insignificant difference in the opinion patterns between controllers who were assessed as level 3 in EPLIS and those who are levels 4 or higher. According to their perceptions, controllers of all proficiency levels feel confident when dealing with non-routine situations in English, even when they do not have a significant level of professional experience.

Level 3 controllers consider their level of proficiency quite sufficient to communicate in non-routine situations in English when phraseology does not suffice and indicate that both native and non-native pilots are generally able to understand their instructions and readback correctly.

Controllers of all levels have faced situations that they could not solve by themselves and had to rely on the more experienced professionals, which seems to corroborate that experience and professional knowledge, although not part of the international language policy, are part of the radiocommunication construct.

References

Alderson, J.C. (2008). *Final report on a survey of aviation English tests*. www.ealta.eu.org/guidelines.htm [Accessed on 3.12.2008].

Bachman, L.F./ A.S. Palmer (1996), *Language testing in practice*. Oxford University Press: Oxford.

Chapelle, C.A./ M.E. Enright/ J. Jamieson (eds.), (2008), *Building a validity argument for the Test of English as a Foreign Language.* London: Routledge.

Chapelle, C.A. (2021), *Argument-based validation in testing and assessmen*t. Sage: Thousand Oaks, CA.

Creswell, J.W. / V.L. Plano Clark (2011), *Designing and Conducting Mixed Methods Research*. 2nd Edition, Sage Publications: Los Angeles.

Creswell, J.W. / L.V. Plano Clark (2018), *Designing and Conducting Mixed Methods Research* (3rd ed.). Sage Publications: Los Angeles.

Departamento de Controle do Espaço Aéreo. Publicações (2018), *Manual de implementação dos requisitos de proficiência em inglês* (MCA 37–225).

Douglas, D. (2014), *Nobody seems to speak English here today: Enhancing assessment and training in aviation English*, (in:) "Iranian Journal of Language Teaching Research" 2(2), 1–12.

Estival, D. / C. Farris/ B. Molesworth (2016), *Aviation English: A lingua franca for pilots and air traffic controllers*. Routledge: Oxford.

ICAO (2010), *Manual on the Implementation of ICAO Language Proficiency Requirements* (2nd ed.): International Civil Aviation Organization: Montreal.

Kane, M. (2006), *Validation*, (in:) R. Brennen (ed.), "Educational Measurement" (4th Edition), Greenwood Publishing: Westport, CT, 17–64.

Kane, M.E. (2013), *Validating the interpretations and uses of test scores*, (in:) "Journal of Educational Measurement" 50 (1), 1–73.

Kim, H. (2012), *Exploring the Construct of Aviation Communication: A Critique of the ICAO Language Proficiency Policy* (Unpublished doctoral dissertation). University of Melbourne, Department of Linguistics and Applied Linguistics.

Kim, H./ C. Elder (2009), *Understanding aviation English as a lingua franca: Perceptions of Korean aviation personnel*, (in:) "Australian Review of Applied Linguistics" 32(3), 23.21–23.17.

Knoch, U. (2014), *Collaborating with ESP stakeholders in rating scale validation: The case of the ICAO rating scale*, (in:) "Spaan Fellow Working Papers in Second or Foreign Language Assessment" vol. 7, 21–46.

Knoch, U./ S. Macqueen (2020), *Assessing English for Professional Purposes*. Routledge: New York.

Kruskal, W.H. / W.A. Wallis (1952), *Use of ranks in one-criterion variance analysis*, (in:) "Journal of the American Statistical Association", 23–27.

Likert, R. (1932), *A technique for the measurement of attitudes*, (in:) "Archives of Psychology", 4–6.

Mann, H.B. / D.R. Whitney (1947), *On a Test of Whether One of Two Random Variables Is Stochastically Larger than the Other,* (in:) "Annals of Mathematical Statistics", 17–19.

McNamara, T./ C. Roever (2006), *Language testing: The social dimension*. Blackwell Publishing: Malden, MA.

McNamara, T. (1996), *Measuring second language performance.* Longman: London.

Pallant, J. (2013), *SPSS Survival Manual: A Step-by-step Guide to Data Analysis using SPSS version 15* (3rd ed.). McGraw-Hill: New York.

Toulmin, S.E. (2003), *The uses of argument*, Cambridge University Press: Cambridge.