A Validation Study on PEPEC Based on Content-related Evidence

Lixin Dai¹,², Jianwen Zhang² and Bing Hao³

¹School of Civil Aviation, Shenyang Aerospace University, Shenyang, Liaoning Province, China
²CAAC Southwest Regional Administration, Chengdu, Sichuan Province, China
³Air China, Beijing, China

*Corresponding author e-mail: dailx55988@sau.edu.cn

Abstract. For the purpose of improving the overall flight English proficiency of Chinese pilots to ensure civil aviation flight safety, PEPEC expert group has conducted a series of researches in 2020 to optimize PEPEC test system. This study, employing ICAO rating scale and open-ended interview, validated PEPEC 900 Sentence Recitation optimized tasks administered in August 2020 on the basis of content-related evidence theory. The results indicate that: 1) test tasks adequately represented the proposed constructs; 2) the objectives defined in each combination task were generally consistent with the test task objectives. Considering the evidence listed above, the conclusion of the study is that PEPEC 900 Sentence Recitation optimized tasks are of high validity in terms of content-related evidence.

Keywords: PEPEC, Civil Aviation, Flight Safety, Validation

1. Introduction
PEPEC is a language test system developed by the Aviation English Expert Group of the Civil Aviation Administration of China (CAAC) in 2009 in accordance with the requirements of the International Civil Aviation Organization (ICAO) for evaluating the level of pilots’ aviation English. In order to improve the overall flight English proficiency of Chinese pilots, the PEPEC expert group has conducted a series of seminars and researches in airlines, aviation colleges and flight schools to improve the construction of PEPEC tasks to meet ICAO language proficiency requirement (LPRs). Based on the validity theory of language testing, this study uses the ICAO raring scale, and open-ended interview to collect evidence on the validation of the 900 Sentence Recitation tasks. The purpose of this study is to justify the assessment use and improve PEPEC system and thus to avoid potential flight safety hazards caused by language factors.

2. Background of the Study
Concerning the international flight incidents and accidents, it is well established that certain air crashes are caused by communication factors [1][2]. Pilot aviation English is part of English for
Special Purpose (ESP), which is composed of land and air call terminology (phraseology) and plain English (Plain English). The International Civil Aviation Organization (ICAO) considers the test of aviation English as one of the three elements to improve the quality of aviation communication[3]. In 2003, ICAO introduced Language Proficiency Requirements (LPRs) for pilots and air controllers worldwide, and later issued the ICAO Language Proficiency Requirements Implementation Manual with a language proficiency scale as a measure of pilot language proficiency. In 2008, ICAO formally introduced the requirement for pilots and air controllers operating on international routes to have their industry licenses endorsed with ICAO Aviation English Language Proficiency Level 4 or higher. In response, national civil aviation authorities and some commercial organizations opened the development and operation of the ICAO Aviation English Language Test. The relevant test is classified as a high-risk level test [3].

China’s PEPEC is a pilot language proficiency test system based on the language proficiency assessment criteria in Doc 9835, and is the only and influential aviation English level test system for flight personnel in China. Aviation flight English is narrow in scope, but has a huge impact, and related research is very limited. In terms of the validity of the test, although there are scholars at home and abroad who have conducted research on general English from different perspectives, there are few research practices in aviation English testing, and there is a lack of research on the validity of the PEPEC test from the perspective of content arguments.

This study takes the content arguments proposed by Bachman (1991) as the rationale, takes the 900-sentence recitation tasks as an example, uses the ICAO LPRs evaluation scale to compare and analyze the optimized combination of 900-sentence test results, and uses open-ended interviews and other forms to collect content evidence and verify its construct validity, in order to improve the construction of the PEPEC question bank and make it better serve civil aviation flight personnel.

3. Theoretical Framework
It is widely accepted in the testing community that the key to determining the quality of language testing and evaluation is validity [4][5][6][7][8]. Bachman (1991) proposed a “test usefulness” perspective, noting its six characteristics: reliability, construct validity, veracity, interactivity, impact, and usefulness. From the 1990s to the present, validity is generally considered as a holistic concept[4][7], which Messick (1989: 7) defines as “the extent to which evidence and theory are based on inferences made or actions taken on the basis of test scores or other forms of assessment”. The adequate and appropriate interpretation and use of test scores is closely related to the holistic view of validity. The holistic view of validity states that

1) validity is an argument about the adequacy and rationality of the interpretation and use of test scores.
2) validity is evidence, and the validity validation process is the process by which test users give evidence about its rationality.

Evidence of validity comes from the following areas: test content, internal structure of the test, response process, relationship between test scores and other external variables, and consequences of the test [9].

Bachman (1991) defines this as content relevance and coverage, validity correlation, and test fairness. In this regard, content relevance refers to the specification of the relevant test performance behavior (i.e., relevant competencies) in relation to the corresponding task [7], and content coverage refers to “the extent to which the test task is representative of all aspects of the competencies being measured” [4]. In other words, content correlation is the correlation between the proposed test subject’s ability constructs and the definition of the test task and the detailed indicator description of the test task. Validity scale correlation validity refers to the degree of correlation between test scores and validity criteria. In other words, the validity of the test content (task and description) can be examined by comparing the arguments of the test scales to the subjects’ use of the test tasks. In addition, test content and coverage constructs can be obtained from the knowledge of the test users (test system developers, examiners, scorers, and test takers) about the test.
Bachman’s (1991) theory of test “usefulness” suggests that authenticity refers to the degree to which a test task is consistent with the characteristics of the task in which the target language is used. The degree to which a test task is consistent with the features of the target language use task. In other words, the consistency of a test task with the intended construct is one of the indications of test authenticity, which can be assessed by comparing the content of the test task with the language features used in the real context of the participant, and this evidence can also be obtained from the participant’s knowledge of the test.

4. Research Methodology

4.1. PEPEC Test and ICAO LPRs Rating Scales

| PEPEC | Test Subjects | Test contents | Evaluation Objective | Test Criteria |
|-------|---------------|---------------|----------------------|--------------|
|       | Civil aviation pilots, flight trainees, other aviation professionals who need to meet ICAO LPRs | Part 1: Listening comprehension (20 questions, multiple choice; 15 minutes), including recorded conversations between controllers and pilots in conventional and non-conventional land and air communication situations. The test is designed to test the test taker’s ability to obtain verbal information. Part 2: 900 sentence recitation (LEVEL 3 with 8 short sentences, 5 minutes; LEVEL 4 with 12 short sentences, 10 minutes; LEVEL 5 with 20 sentences, 15 minutes), which requires recitation of ICAO standard terminology or general English for non-routine and emergency situations. The test is to test the ability to hear and understand correctly the terms used in conventional and non-conventional air and land calls. Part 3: Simulated Land-Air Call (15 minutes), which simulates the interaction between pilots and controllers using ICAO standard terminology and general English during routine and non-routine land-air calls. The test examines the test taker's overall ability to use language in a simulated real-world working environment. Part 4: Story retelling (LEVEL 5 and above, 10 minutes) requires the retelling of a story related to the aviation profession, approximately 60-100 seconds in length and 100-120 words per minute (listen twice). The test will test the ability to express the content of the test, the ability to organize the text and the ability to use language. Part 5: Oral examination (semi-direct examination based on computer video, 15 minutes), in four stages: preparation stage, preliminary rating stage, rating stage and closing stage. The examiner asks the candidate questions on issues related to the aviation profession and tests the candidate's overall language skills. | Assessment of the candidate's level of aviation English proficiency (Levels 3, 4 and 5) | ICAO LPRs rating scales |

The PEPEC (Pilots’ English Proficiency Examination of CAAC - the English Proficiency Examination System for Civil Aviation Personnel in China) is a web-based and computer-based “criterion-related reference” system for testing English proficiency in aviation (see Table). The content, scope, and standards of the PEPEC language test are based on ICAO Document 9835 [10][11], and the ICAO Handbook (ICAO, 2004) provides a scale as a reference for assessing the test.
taker’s language proficiency. The scale is a descriptive scale with six levels of language proficiency from 1 to 6, namely sub-elementary (level 1), elementary (level 2), sub-operational (level 3), operational (level 4), extended (level 5), and expert (level 6). Working level (level 4) is the lowest level of language proficiency that pilots are expected to achieve. The scale breaks down the six levels from the six structures of language proficiency - pronunciation; structure; vocabulary; fluency; comprehension; and interactions. Refine the language skills required of the participant. Civil aviation pilots, flight trainees, and other aviation professionals who need to meet ICAO LPRs. See the table below.

4.2. Research Questions and Steps
This study involves PEPEC Part 2 and intends to answer the following two questions: 1) Does the proposed test constructs of the optimized PEPEC 900-sentence recitation combination tasks meet ICAO test criteria? 2) Does the optimized PEPEC 900-sentence recitation combination task authenticity feature meet users’ expectations? The participants in this study included: 10 PEPEC language test experts from the Civil Aviation Administration of China Southwest Region (leader of the expert group), Civil Aviation Administration of China East Region, Air China, China Southern Airlines, China Eastern Airlines, Civil Aviation University of China, and Shenyang aerospace University. Eight of them were experts in the flight industry and two were experts in linguistics. The participants in this study included 100 active pilots, including captains and co-pilots, aged 20-50, from China Southern Airlines Southwest, Air China Southwest, and Tibet Airlines; and 15 students and teachers of flight technology from Shenyang aerospace University.

To answer the above two questions, this study adopted the research method of comparative analysis between groups and open-ended interviews. In order to assess the validity of the optimized PEPEC content concept, the expert group developed four task combinations of “900 sentence recitation” based on the ICAO definition of aviation English proficiency and the PEPEC test content requirements, combined with the LPRs (ICAO, 2004) rating scale. Each combination task covers radiotelephony communication terminology at levels of difficulty 1-6 as described in ICAO handbook. Each combination consists of 12 questions with the following distribution of difficulty: combination 1: 4 sentences (level 3) + 4 sentences (level 4), which is higher than the difficulty of the test pool; combination 2: 4 sentences (level 3) + 4 sentences (level 3) + 2 sentences (level 3) + 2 sentences (level 4), which is equal to the difficulty of the existing questions; combination 3: 4 sentences (level 2) + 4 sentences (level 3) + 4 sentences (level 3) with a difficulty level slightly lower than the difficulty of the existing questions; Combination 4: consists of 4 sentences (Level 2) + 4 sentences (Level 2) + 2 sentences (Level 2) + 2 sentences (Level 3) with a difficulty level lower than the difficulty of the existing questions. Each question is worth 5 points, for a total of 60 points. Level 6: 54 points or more; Level 5: 48 to 53 points; Level 4: 42 to 47 points; Level 3: 36 to 41 points; Level 2: less than 35 points. Subjects were divided into 4 groups according to their age (50+, 40-49, 30-39, and under 30). The selected utterances for the 4 combinations were basically at the level of 3-4.

Subjects completed the recitation of the 4 task combinations according to the test instructions, and members of the expert panel (divided into 4 groups) rated the subjects’ accuracy of the recitations based on ICAO LPRs rating scales.

5. Results and Discussion
The statistical results of the survey showed that the optimized PEPEC 900-sentence recitation task type conformed to the ICAO LPRs range and adequately represented the proposed constructs, and that the test tasks examined detailed indicators of the proposed constructs. The data showed that, in the absence of any test preparation, the results of the combination test for subjects of different ages in the three companies were consistent with the proposed combination test difficulty. For example, the test results for combinations 2, 3, and 4 were generally at the ICAO LPRs Level 4 level, indicating that the subjects’ performance was adequately reflected by the description of the content of the questions, in other words, the proposed test content concept was generally consistent with the requirements of the
described test tasks. The slightly lower test scores for combination 1 relative to combinations 2, 3,4 indicate that the test result arguments are consistent with the relevance of the task description for that combination. The overall distribution of the subjects’ scores indicated that the optimized PEPEC 900-sentence recitation question constructs met the validity test criteria of the ICAO LPRs. The results of the different task combinations for different age groups suggest that the optimized 900-sentence combination tasks can be used by airlines to test the effectiveness of training for specific age groups.

Evidence for “content relevance and content coverage,” “task feature construct rationality,” and “content authenticity” was also collected from open-ended post-test interviews with pilots from the three airlines participating in this experiment, as well as random interviews with faculty and students from Shenyang Aerospace University. The evidence for “content relevance and content coverage,” “task feature conceptualization rationality,” and “content authenticity” was also collected from open-ended post-test interviews with three airline pilots who participated in this experiment and random interviews with flight technology students and faculty at Shenyang Aerospace University. The results showed that 98.75% of the respondents believed that “the content of the questions is highly relevant to the professional working environment.”, 97.53% of the respondents thought that “the optimized recitation contents are widely used in pilots’ work and highly relevant to aviation communication scenarios.” The questions reflect the professional character of the test and the rationality of the task characteristics. In terms of agreement with the “different combinations of difficulty”, test users (test takers and examiners) generally agreed that the four combinations covered different levels of radiotelephony communication utterances, and that the different levels of complexity of the sentences within the combinations reflected the professional aviation language level of the language users, indicating that the optimized combinations adequately reflected the test content. The optimized test combinations fully reflect the content of the test in terms of the concept of competence coverage and the authenticity of the content.

To sum up, the PEPEC optimized 900-sentence combination tasks meets the ICAO test criteria, and its task authenticity features meet the test users’ expectations. The study indicates that PEPEC can develop and conduct reform tasks according to the current combinations; training and teaching of aviation English in airlines and aviation colleges should tend to refer to the ICAO LPRs scoring standards and assess the learners’ professional radiotelephony in conventional and non-conventional situations in the form of optimized 900-sentence combinations at different levels, so that training and teaching can be more targeted. Although the data collected in this study involves only some airlines and one aviation college, and cannot summarize the whole industry, the expert group only hopes to provide a way of thinking for the PEPEC test research, so that the civil aviation industry can understand the current situation of PEPEC development and reform measures.

6. Conclusion
Charles Alderson [13] [14] argues that the paucity and inadequacy of research on language factors in aviation may lead to unreliable, poorly validated, and inappropriate language tests, and therefore to an increase in fatalities caused by incorrect communication between pilots and air traffic controllers. The current state of affairs is that aviation English language testing is still in the process of development, and relevant researches are very limited. The importance of PEPEC as the authoritative test system for aviation English testing in China is evident in its test validity. This paper reports the results of the recent research on the optimization of 900-sentence recitation tasks, one of the PEPEC question bank construction projects implemented by the PEPEC Aviation English Expert Group in China, and makes constructive suggestions and points out the limitations of the research. The aim is to keep pilots of airlines, flight technology students of aviation colleges, flight trainees and other aviation professionals abreast of the current status of PEPEC development, so that PEPEC can better serve civil aviation professionals, avoid flight accidents caused by language factors and ensure flight safety.
Acknowledgment
The authors would like to thank PEPEC expert group; civil aviation administrators, pilots and staff in China Southern Airlines, China Southwest Airlines and Tibet Airlines; teachers and student pilots from Shenyang Aerospace University for their great contribution and valuable comments on this study.

References
[1] D. Estiva, C. Farris, and B. Molesworth, A lingua franca for pilots and air traffic controllers. Routledge Taylor & Francis Group, London and New York. 2016.
[2] B. Molesworth, et al. (2014) “The effect on recognition memory of noise cancelling headphones in a noisy environment with native and non-native speakers” Noise and Health, 16 (17),240-47, 2014
[3] ICAO. Manual on the implementation of ICAO language proficiency requirements. Doc. 9835. 1st edition. International Civil Aviation Organization. 6.1.3, 2004.
[4] S. Messick, Validity, in Educational Measurement(3rd. ed.). New York: Macmillan, 1989.
[5] C. Weir, Language Testing and Validation: An Evidence-Based Approach. Palgrave: Macmillan, 2005.
[6] C. Alderson et al., Language Test Construction and Evaluation. Cambridge University Press, 1995.
[7] L. Bachman, What Does Language Testing Have to Offer? TESOL Quarterly 25: 671–704. 1991.
[8] R. Yan, Assessing the English Language Proficiency of International Aviation Staff. In The Companion to Language Assessment Vol.484-489, 2014.
[9] APA. Standards for Educational and Psychological Testing. Washington, DC, 1999.
[10] Manual on the Implementation of ICAO Language Proficiency Requirements DOC 9835 International Civil Aviation Organization, Chapter2, Chapter 3.(2004).
[11] Manual on the Implementation of ICAO Language Proficiency Requirements DOC 9835 International Civil Aviation Organization, Chapter2, Chapter 3.(2010).
[12] C. Alderson, The politics of aviation English testing. Language Assessment Quarterly, 8(4), 386–403, 2011.
[13] C. Alderson, Air safety, language assessment policy and policy implementation: The case of aviation English. Annual Review of Applied Linguistics, 29, 168–187., 2009.
[14] C. Alderson, Final report on a survey of aviation English tests. http://www.ealta.eu.org/guidelines.htm, 2009.