© 2023 University of Warsaw. This is an open access article under the **CC BY-NC-ND** license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Applied Linguistics Papers 27/1: 2023, 14–39

Investigating the Construct of a Listening Test to Assess Pilots' Comprehension: a Step-by-Step Project for Test Developers

Angela Carolina de Moraes GARCIA
Carleton University
E-mail: angelagarcia@cmail.carleton.ca,

Abstract: Pilots and air traffic controllers must demonstrate their ability to listen and speak the language used in radiotelephony communications demonstrated by completing a language test. In this context, it is crucial to assess both interactive listening, when listening occurs together with speaking, and listening in isolation, when there is no speaking or interaction. The purpose of assessing listening in isolation is to reduce the influence of skills that are not relevant to the construct, that is, to minimize construct irrelevant variance (S. Messick 1989). This article describes a project that can be followed by test developers to address the initial step in the development of a test to assess pilots' listening in isolation: the construct definition. The project is framed within an interactionalist perspective wherein a test construct is defined based on a combination of the abilities that those taking the test should have and the tasks that they should be able to perform (L. Bachman 2007). It is also informed by the work of L. Bachman/ A. Palmer (2010) and the framework proposed by U. Knock/ S. Macqueen (2020) for the development of language assessments for professional purposes. The project outlined in this article may also be of interest to test developers who wish to investigate different constructs of aeronautical English tests, as well as those involved in the development of other types of language assessments for professional purposes.

Keywords: language testing, language assessment for professional purposes, ICAO language proficiency requirements, listening comprehension, construct definition

Introduction

This paper illustrates one possible way to define the construct of a language test in the context of assessing languages for specific purposes (LSP), or more specifically, in the context of language assessment for professional purposes (LAPP). U. Knoch/ S. Macqueen (2020: 3) define LAPP as "any assessment process, carried out by and for invested parties, which is used to determine a person's ability to understand and/or use the language of a professionally-oriented domain to a specified or necessary level." Although the project outlined in this paper aims to illustrate the steps that could be taken to investigate the construct of a listening test in the context of aeronautical English testing, it may also be of interest to test developers who wish to investigate different constructs of aeronautical English tests, as well as those involved in the development of other types of language assessments for professional purposes. This paper also intends to bridge the gap between theory and practice by outlining practical steps that are informed by the latest research and language assessment theory.

The term aeronautical English refers to the English used by pilots and air traffic controllers (ATCOs) in their radiotelephony (RT) communications (P. Tosqui-Lucks/ A. Silva 2020). According to the International Civil Aviation Organization (ICAO)'s language proficiency requirements (LPRs), pilots and ATCOs' ability to speak and understand the English used in RT communications must be assessed. ICAO developed a rating scale to be used in such assessments, the ICAO rating scale, which includes six areas of evaluation: pronunciation, structure, vocabulary, fluency, comprehension, and interactions. Moreover, Criterion 3 of the ICAO test design guidelines, developed by the International Civil Aviation English Association (ICAEA), states that "test instruments need to contain tasks dedicated to assessing listening comprehension, separate from tasks designed to assess speaking performance" (ICAEA n.d.a). The proposed project aims to assess listening in isolation, as required by criterion 3. This type of assessment is necessary because it allows listening to be assessed by itself, without much interference from other skills. A central principle in language testing is to minimize construct irrelevant variance by reducing the influence of skills that are not relevant to the construct (S. Messick 1989). Although it is necessary to assess pilots and air traffic controllers' listening comprehension in isolation, this should be done in addition to and not to the exclusion of the assessment of interactive listening during the speaking test, as, most of the time, pilots and ATCOs need to listen and to interact with each other simultaneously.

In order to develop a listening test to assess pilots' listening in isolation, it is necessary to have a clear definition of the construct to be assessed. An important consideration is that listening happens inside our minds, so it cannot be assessed directly, as speaking and writing (J. Field 2019). L. Harding (2015: 123) argues that, in language testing, listening is "still a very under-represented skill". In the context of aeronautical English assessment, a fairly new field of research, this problem is even more prominent. The process for creating a high stakes test, such as a listening test for pilots and ATCOs, is complex, time-consuming and requires a great effort from the needs analysis to the operationalization of the test, including the definition of the construct, the writing of test specifications, the development of tasks, the trial of test items, the development of the scoring criteria, and the validation of the test. The goal of the proposed project is to address the initial stage of the development of a test to assess pilots' listening comprehension in isolation: the construct definition.

Thus, the proposed overarching question to be addressed in the proposed project is:

What should be the construct of an aeronautical English listening test for pilots?

A similar question could be formulated for ATCOs and the project described here could be adapted for ATCOs rather than pilots. However, in keeping with the ICAO test design guidelines that "separate test instruments need to be designed for pilots and air traffic controllers" (Criterion 2, ICAEA n.d.b), different listening tests should be developed for pilots and ATCOs. In order to address this question, a *needs analysis* should be conducted. A needs analysis is an investigation of the target language use (TLU) domain (i.e., "a specific setting outside the test itself that requires the test taker to perform language use tasks" [L. F. Bachman/ A.S. Palmer 2010: 60]), and is an essential step in the development of an assessment of LSP, especially in the development of LAPPs. According to U. Knoch/ S. Macqueen (2020: 83), a careful needs analysis helps to "increase the

trustworthiness of an assessment instrument". In conducting this needs analysis, "the contributions of policy, practices, selected social theories, empirical research, and multiple stakeholders" (A. Monteiro/ J. Fox 2022: 170) should be taken into consideration.

To conclude this introduction, some questions are suggested. In the next section, the suggested methodology is explained, including an explanation about the theoretical foundation that could underpin such project. The conclusion includes suggestions for future steps to be taken, some limitations of the proposed project, and some final considerations. The appendices include provisional interview, questionnaire and focus group questions that could be used for the project.

1. Specific questions that test developers may address

As A. Monteiro/ J. Fox (2022: 194) argue, "in multicultural professional contexts in which participants use ELF [English as a Lingua Franca] alongside workplace-specific terminology, such as international radiotelephony communications in aviation, (…) test contexts and constructs should be defined based on characteristics of the TLU domain anchored in the perspectives and accounts of domain stakeholders". With this objective in mind, some specific test development questions were formulated as well as a methodology to address them. The suggested questions are listed in Table 1 and have been formulated within an interactionalist perspective wherein a test construct is defined based on a combination of the abilities that those taking the test should have and the tasks that they should be able to perform (L. Bachman 2007). As L. Bachman (2007: 42) explains, the interactionalist perspective "views the construct we assess not as an attribute of either the individual language users or of the context, but as jointly co-constructed and residing in the interactions that constitute language use". Defining the construct based on the interaction between both abilities and tasks has not only been recommended (e.g., C. Chapelle 1998; M. Chalhoub-Deville 2003), but it has been argued to be the most appropriate approach in the case of listening tests where the listening performance is a result of the underlying knowledge and ability, the situational factors, and the interaction between them (G. Buck 2001). This approach is especially useful in the context of LAPPs because of the importance of the professional domain to the language assessment. As U. Knoch/ S. Macqueen (2020: 63) point out, "the goal of LAPP is to extrapolate an ability classification based on a brief performance sample in a relatively contrived context to other contexts of use in which professional knowledge is central". Thus, I suggest following what L. Bachman (2007) calls a moderate interactionalist approach to construct definition: the "an ability – in language user – in context" perspective suggested by M. Chalhoub-Deville (2003). This approach highlights the importance of considering context when investigating test constructs because, according to M. Chalhoub-Deville (2003: 369), "individual ability and contextual facets interact in ways that change them both".

Specific test development questions (STDQ)	
1	What are the available resources to develop and administer a test to assess pilots' listening comprehension in isolation, and what are the expectations of the organization which is developing the test or to which the test is being developed?
2	According to the ICAO regulations and guidelines, what types of skills, knowledges and processes should be assessed in a test that aims to assess pilots' listening in isolation, and what are the TLU domain tasks and their characteristics?
3	According to the academic literature and research, what types of skills, knowledge and processes are needed in the TLU domain and thus should be assessed in a test that aims to assess pilots' listening in isolation, and what are the characteristics of the TLU domain tasks?
4	What types of skills, knowledges, and processes are assessed in a recognized listening test for pilots, and what TLU domain tasks are represented in it?
5	According to key stakeholders (e.g., pilots and/or ATCOs, raters, researchers), what types of skills, knowledge and processes are needed in the TLU domain and should be assessed in a test that aims to assess pilots' listening in isolation, and what are the characteristics of the TLU domain tasks?
6	Based on the accounts of aeronautical English researchers, how could the construct defined in the draft *design statement* be refined?

Table 1. Specific test development questions (STDQ).

The goal of STDQ 1 is to find out what resources are available and to learn about the needs and expectations of the organization that is requiring the development of the test. STDQ 2 aims to investigate the construct according to the ICAO regulations and guidelines (e.g., ICAO 2010; ICAO 2020), including the ICAO rating scale, and the ICAO test design guidelines, developed by ICAEA (ICAEA n.d.c)[1]. I believe a detailed inspection of the policy and the guidelines is a good starting point for the development of a listening test. As C. Moder/ G. Halleck (2021: 82) suggest, one of the first step to be taken to develop a LSP test is "to consult relevant information provided by government agencies and by professional groups charged with the language training of the professionals in the target domain". After that, I suggest carrying out secondary research to investigate how academic research and literature can contribute to the definition of the construct of a listening test for pilots, by conducting a thorough review of the literature (STDQ 3). STDQ 4 aims to investigate the construct of an existing listening test for pilots or ATCOs. At the time of writing this paper, the only test endorsed by ICAO[2] was the ELPAC for ATCOs. The ELPAC, which stands for *English Language Proficiency for Aeronautical Communication*, was developed by the European Organisation for the Safety of Air Navigation (EUROCONTROL), in partnership with other institutions. Looking into the construct of an existing listening test may valuably inform the decisions to be made in relation to what

[1] At the time this paper was written, a group of experts which ICAO invited were revising the ICAO test design guidelines in order for ICAO to have it published as a handbook.
[2] See https://www4.icao.int/aelts/Home/RecognizedTest

construct should be measured in a test to be developed. The goal of STDQ 5 is to investigate the perceptions of key stakeholder such as pilots and/or ATCOs on what the construct of a listening test should be, similarly to what A. Garcia/ J. Fox (2020) have done. It is pivotal to investigate the perceptions of domain experts in relation to the construct to be assessed as "language and communication may mean one thing to linguistically oriented professionals and another to gatekeepers within a professional community" (S. Jacoby/ T. McNamara 1999: 236). Finally, STDQ 6 aims to investigate the perceptions of aeronautical English researchers about the construct that should have been defined based on the data gathered in the study. The purpose is to have the researchers evaluate the definition of the construct so it can be improved. If access to researchers is difficult, you may consider inviting other key stakeholders to evaluate the defined construct (e.g. aeronautical English test raters, teachers, etc.). Figure 1 shows an overview of the proposed project, including the specific test development questions each strand addresses.

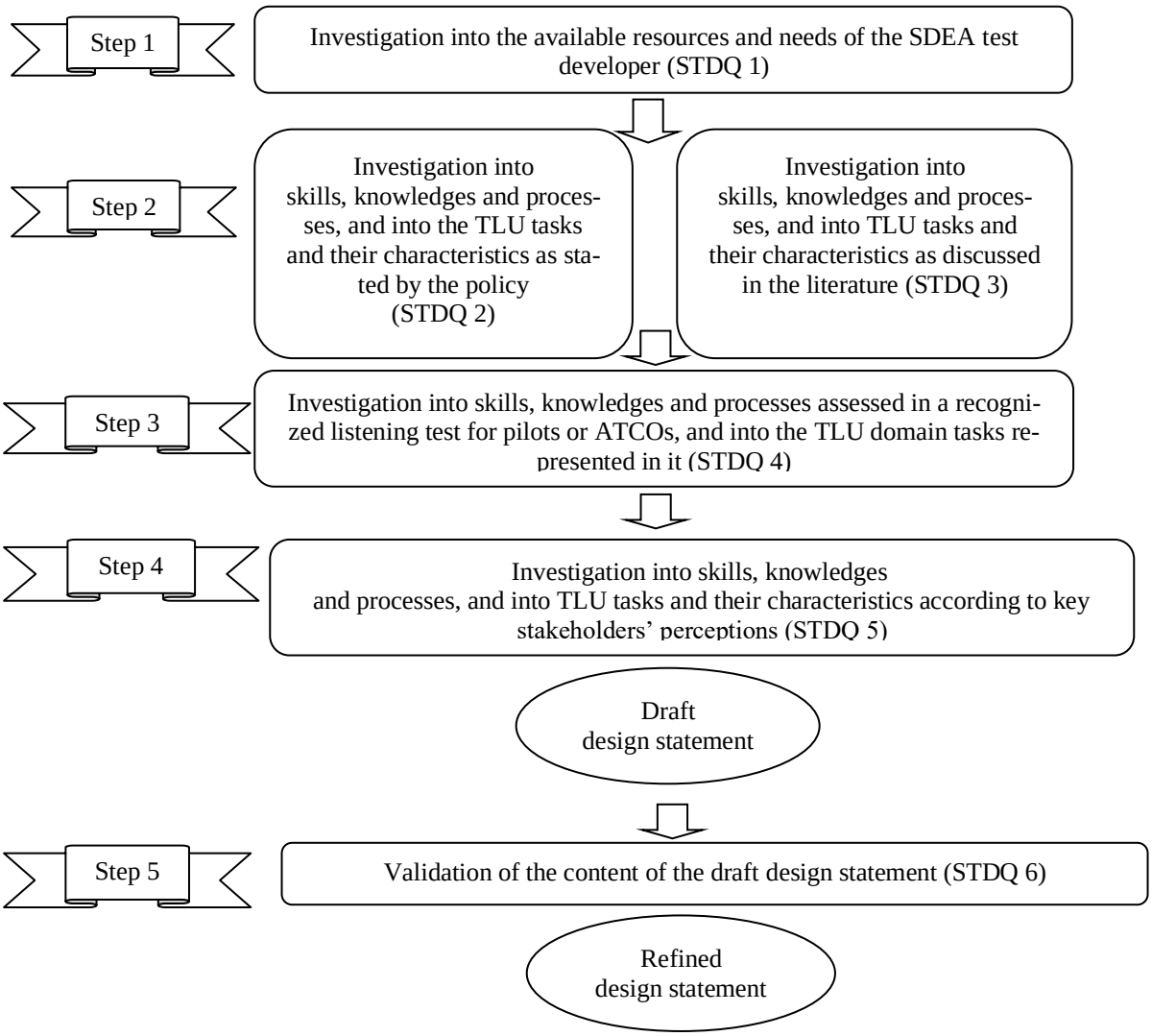

Figure 1. Overview of proposed project.

2. Proposed methodology

In the proposed project, I follow a pragmatic approach, which advocates the use of "a straightforward, 'to-the-point' research methodology that offers a great deal of practical help" (Z. Dornyei 2007: 18). However, good research must follow a sound theoretical framework, as well as a strong methodological approach. Therefore, I recommend using as the conceptual foundation for the proposed project the socio-cognitive theory developed by C. Weir (2005). The socio-cognitive approach to testing, as explained by M. Milanovic/ C. Weir (2013: x), "seeks to take account of both the aspects of cognition, related to the mental processes the individual needs to engage in order to address a task, and the features of language use in context that affect the ways in which a task is addressed". I also suggest following the recommendations to develop language assessments for professional purposes given by U. Knoch/ S. Macqueen (2020). Other useful works which may underpin the project include L. Bachman/ A. Palmer (2010) and D. Douglas (2000). In relation to the methodological approach, needs analyses usually draw on multiple sources and methods (U. Knoch/ S. Macqueen 2020). Therefore, I recommend conducting a multiple methods study, in which the strengths of both qualitative and quantitative approaches can be combined in a complementary way (Z. Dornyei 2007), as gathering both qualitative and quantitative data helps to develop a better understanding of the research problem than using only one approach (J. Creswell 2015). U. Knoch / S. Macqueen (2020: 96) explain that "drawing on multiple data sources to triangulate the results increases the credibility of the conclusions drawn based on the needs analysis". Thus, the proposed project uses three types of triangulation, as explained by J. Brown / T. Rodgers (2002): data triangulation, by using multiple sources of information (data from policy, literature, and different stakeholders – pilots/ATCOs, test developers, researchers); theory triangulation, by using multiple theoretical frameworks (D. Douglas 2000, M. Chalhoub-Deville 2003, L. Bachman/ A. Palmer 2010, U. Knoch/ S. Macqueen 2020), and; methodological triangulation, by using different procedures to collect data (literature review, interviews, focus group, questionnaires, document analysis).

2.1 The theoretical framework and its relationship with the proposed study

U. Knoch/ S. Macqueen (2020) propose a socially-oriented theory of construct for LAPP, a model of needs analysis, procedures for turning the results of a needs analysis into a test blueprint and specifications, as well as a framework to be used for validation of LAPP. For a better understanding of this proposed study, it is important to, firstly, define the four dimensions of construct proposed by U. Knoch/ S. Macqueen (2020): the stated construct, the operationalized construct, the theoretical construct, and the perceived construct.

- The *stated construct* refers to what is publicly claimed to be assessed on a test. Information about the stated construct can be found on the test's website (description of the test, sample test), on the policy, the rating scale, etc.;
- The *operationalized construct* refers to what is really being assessed during the actual test. For researchers to gather information about the operationalized construct, they would have to have access to the actual performances of test takers (e.g., their responses) or to their behaviour during the test, in order to investigate what they are, for example, thinking or doing during the test;

- Another dimension of construct, the *theoretical construct*, is unobservable and refers to the theory on which the assessment is based. U. Knoch/ S. Macqueen (2020: 40) explain that "in LAPP, this is typically a language proficiency, skill or ability that is assumed to underlie communication in an actual, specific world of work". The theoretical construct might be explicitly stated or not. Whatever the case may be, information about it has likely guided the development and design of the test, including the test procedures;
- Lastly, the *perceived construct* refers to how participants in the testing process (test takers, raters, policy makers, teachers, etc.) understand the construct (what they believe the test is testing).

Table 2 shows the spheres of construct related to the STDQ 2 to 5:

STDQ	Sphere of construct
2	*Stated and theoretical construct*
3	*Theoretical construct*
4	*Stated and perceived constructs*
5	*Perceived construct*

Table 2. Sphere of construct related to STDQ 2 to 5 (U. Knoch/ S. Macqueen 2020).

Thus, STDQ 2 addresses the stated construct in the ICAO regulations and guidelines, and also the theoretical construct explained especially in the ICAO test design guidelines. STDQ 3 aims to investigate the theory on which the listening test for pilots should be based. STDQ 4 investigates how the construct of a recognized test is stated on their website and publicly available documents, and how it is perceived by its test developer. Finally, STDQ 5 aims to investigate how key stakeholders (pilots and/or ATCOs) perceive the construct to be assessed. The sphere of operationalized construct is not included because it is only possible to start looking into it after the test starts being operationalized.

Secondly, U. Knoch/ S. Macqueen (2020) proposed a cycle of development of LAPP. As shown in Figure 1, the stages for the development of LAPP start with the needs analysis and end with the operational use of the test. However, there needs to be regular review, and the work at a certain stage may require that test developers go back to a previous stage. The proposed project addresses the first stage of the LAPP test development cycle, the needs analysis, and part of the second stage, the development of the design statement, as shown by the blue rectangle in Figure 2.

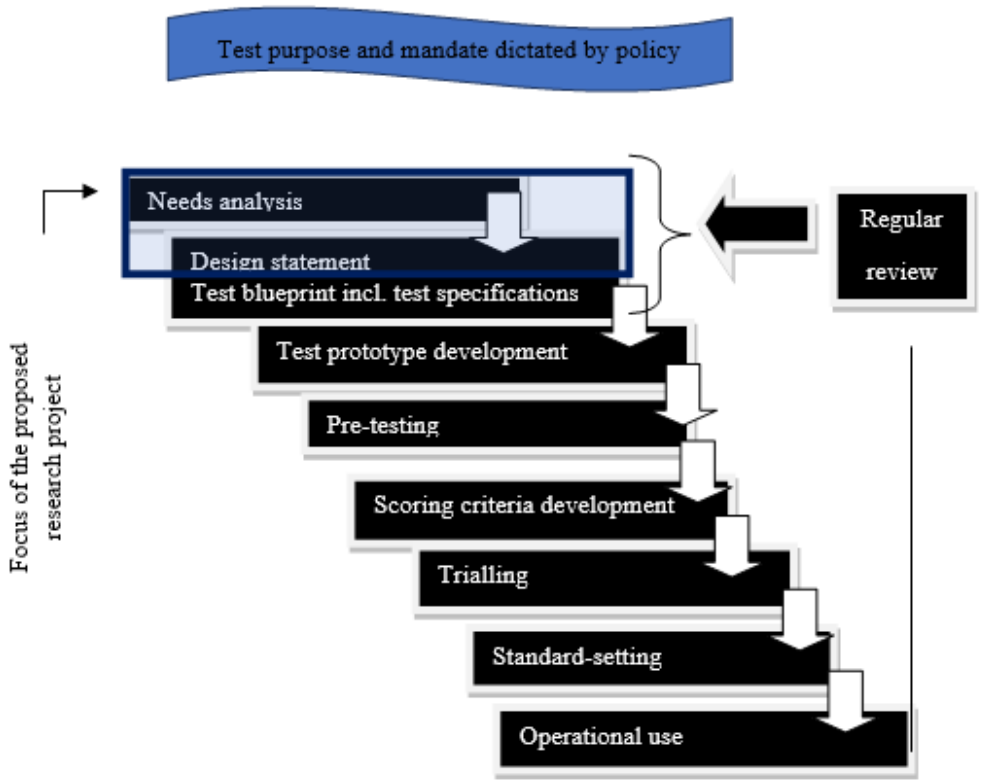

Figure 2. LAPP test development cycle, adapted from U. Knoch/ S. Macqueen (2020: 97).

Thirdly, it is worth considering U. Knoch/ S. Macqueen's (2020) approach to needs analysis. These researchers talk about five areas a needs analysis for developing a LAPP may address: the domain analysis, the means analysis, the policy analysis, the test requirement analysis, and the test taker analysis. The *domain analysis* is a key component in a needs analysis. It requires an empirical analysis of the TLU communication tasks, of the language used to complete these tasks, as well as of the real-world interaction between test takers and tasks. The *means analysis* is an analysis of the available resources for test development, administration and validation. The *policy analysis* is an analysis of the regulations that are relevant to the LAPP. The *test requirement analysis* is an analysis of the requirements for the test, such as test purpose, information about the needs of different stakeholders, and about score reporting and score uses. Finally, the *test taker analysis* is an analysis of the language proficiency of test-takers and of the difficulties they may encounter when communicating in real life, as well as an analysis of the test takers' perceptions of the test, their needs and motivations, and the impact that the test might have on teaching. Table 3 shows the areas of needs analysis addressed by each specific test development question. As U. Knoch/ S. Macqueen (2020) do not mention a kind of needs analysis that aims to investigate the construct by looking at other tests, I have named this kind of analysis *external tests analysis*. U. Knoch/ S. Macqueen (2020) explain that these types of needs analyses do not necessarily need to follow a linear sequence or come before all

stages of assessment design. They follow a cyclical procedure. We can understand them as independent activities that overlap each other.

STDQ	Areas of needs analysis addressed
1	Means analysis Test requirement analysis
2	Policy analysis Test requirement analysis Domain analysis
3	Policy analysis Test requirement analysis Domain analysis Test taker analysis
4	External test analysis
5	Domain analysis Test requirement analysis Test taker analysis

Table 3. The areas of needs analysis for STDQ 1 to 8.

As U. Knoch/ S. Macqueen (2020: 108) point out, "needs analyses generally result in an array of information which is often difficult to consolidate". They suggest, similarly to what L. Bachman/ A. Palmer (2010) had suggested, to start with the development of a draft *design statement* based on the information gathered during the needs analysis. The design statement should contain information to guide the next stages of test development, including a description of the purpose of the test, a description of the test-takers, a description of the TLU, and the definition of the construct to be measured.[3] The information contained in this document should also serve as evidence (or backing) for the warrants in the validation framework. L. Bachman/ A. Palmer (2010) give a more detailed structure for a design statement. They suggest, for example, that a design statement should also include a list of "tasks selected as a basis for developing assessment tasks," as well as a "description of the characteristics of the TLU tasks that have been selected as a basis for assessment tasks" (L. Bachman/ A. Palmer 2010: 270). U. Knoch/ S. Macqueen (2020) suggest that this list of tasks, which they call a *table of TLU tasks*, should be created after the design statement is produced. However, this list (or table) can be incorporated in the design statement, as recommended by L. Bachman/ A. Palmer (2010). Thus, Table 4 shows the information that the design statement should include:

[3] Although other authors (e.g., J. Alderson et al. 1995, R. Green 2017) have recommended that the definition of the construct should be included in the test specifications document, the project explained in this paper follows L. Bachman and A. Palmer's (2010) guidelines. The construct should be defined in the design statement. Future blueprint should include the assessment specifications and the task specifications (including, the construct for each task type).

Structure of Design Statement		
1		Purpose of the test
2		Description of the test takers
3		Definition of the construct
	3.1	Description of the TLU domain
	3.2	Types of skills, knowledges and processes needed in the TLU domain.
	3.3	List of TLU domain tasks selected as a basis for developing assessment tasks
	3.4	Description of the characteristics of the TLU domain tasks

Table 4. Information that might be included in the Design Statement.

The detailed explanation given by L. Bachman/ A. Palmer (2010) about the procedures to be followed to develop a design statement should also be very helpful, as well as the projects they have shared[4]. Although L. Bachman/ A. Palmer (2010) list the sections on the description of the TLU domain, on the tasks, and on the characteristics of TLU tasks separately, I have joined them under the same section because according to the interactionalist approach on construct definition, the construct should be defined based on both traits and context. For section 3.2 of the design statement, I suggest following D. Douglas's (2000) framework of components of specific language ability. For section 3.4, I recommend adapting L. Bachman/ A. Palmer's (2010) framework of test task characteristics, because, as they indicate, their framework may be "useful for describing both TLU tasks and test tasks" (L. Bachman/ A. Palmer 1996: 57).

When designing the interview and questionnaire questions to be used, two issues related to the assessment of LSPs should be taken into consideration. They are *specificity of content* and *inseparability*. As discussed by D. Douglas (2001), these two characteristics of LSP testing, may bring about some theoretical and practical problems[5]. First, the issue of specificity deals with the problem of how specific test tasks should be. For example, should there be a test for all pilots, or a test for airplane pilots and another one for helicopter pilots? If a test was designed for airplane pilots only, there should still be issues because airplane pilots have different flying experiences, they fly different airplanes, and so on. So, the question here again would be *how specific should the test be?* Second, the problem of inseparability handles the understanding that general purposes language tests should not include the assessment of background knowledge, because that would be considered irrelevant for the construct. However, as D. Douglas (2001) argues, in LSP testing, it might not be possible to separate language knowledge from specific purpose background knowledge. For this reason, he advocates that "we must, in testing language for specific purposes, define specific purpose language ability as comprising both language

[4] Available at http://www.oup.com/LAIP.

[5] D. Douglas (2001) also discusses a third problem in assessing LSP: authenticity. In the project explained in this paper, the issue of authenticity does not need to be addressed because test tasks will not yet be developed; the proposed project only gets to the point of listing the TLU tasks and describing their characteristics.

knowledge and background knowledge" (D. Douglas 2001: 50). Thus, we need to ask ourselves *to what extent should the test assess professional knowledge?*

Lastly, although this paper proposes a project to be followed at the beginning of the test development cycle, issues related to validity should be discussed. As L. Bachman/ A. Palmer (2010) argue, the main purpose of language assessments is to collect information for making decisions that will ideally lead to beneficial consequences for stakeholders. Test developers need to be accountable to stakeholders because many tests, including the aeronautical English tests, are high stakes. L. Bachman (2015: 7) explains that high stake tests "have major, life changing consequences for stakeholders, and decision errors (false positive/negatives) are difficult to reverse". Being accountable to stakeholders means that test developers need to be able to justify the use they make of an assessment. In other words, they need to show to their stakeholders that the intended uses of their assessment are justified (L. Bachman/ A. Palmer 2010). Stakeholders are all those involved in or affected by the assessment. In the aeronautical English context, stakeholders can be, to name a few, test takers, test developers, raters, regulators, aeronautical English teachers, airlines, passengers, and society in general. Thus, *validation* and *validity* are central concepts in language assessment. Validation may be defined as "the ongoing process of justifying particular interpretations and uses of test results" (C. Chapelle 1998: 33). Validity is an abstract theoretical term which was defined by S. Messick (1989: 13) as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment". Thus, validation is an activity through which validity is investigated. To guide this process of validation, a conceptual framework is necessary. M. Kane (2002, 2006) presents an argument-based approach to validity. As C. Chapelle et al. (2008: 6) explain, "arguments are used to build a case for a particular conclusion by constructing a chain of reasoning in which the relevance and accuracy of observations and assertions must be established and the links between them need to be justified". An argument-based approach consists of a systematic way to investigate the validity of the interpretations and uses of scores (we should not claim that a test is valid, but we could claim that the interpretations and uses of test scores are valid). As U. Knoch/ S. Macqueen (2020) point out:

> Argument-based approaches to validation rely on specifying a series of inferences, warrants and assumptions associated with score interpretations and uses. Inferences connect a series of claims we make. Underlying each inference, warrants and assumptions are formulated which need to be supported by evidence (also referred to as backing) so we can argue that the inference is supported (U. Knoch/ S. Macqueen 2020: 139).

U. Knoch/ S. Macqueen (2020) propose a validation framework to be used in LAPP. Their validation framework is underpinned by M. Kane's (2002, 2006) work in educational measurement and by language assessment theories (C. Chapelle et al. 2008, L. Bachman/ A. Palmer 2010, U. Knoch/ C. Chapelle 2018). Their framework "is useful in that it provides an overarching framework which is connected by a series of inferences leading from the domain, the assessment materials to the test consequences" (U. Knoch/ S. Macqueen 2020: 164). U. Knoch/ S. Macqueen's (2020) framework includes infor-

mation on both the competences and the context for score interpretation and use, in accordance with the interactionalist perspective to construct definition. The structure of U. Knoch/ S. Macqueen's (2020) validation framework includes seven inferences: domain description, evaluation, generalization, explanation, extrapolation, decisions, and consequences (U. Knoch/ S. Macqueen 2020: 139). Their proposed framework is not fixed; it should be adapted to each particular testing context. The domain description inference (claim, warrants and assumptions) proposed by U. Knoch/ S. Macqueen (2020) highlights the importance of selecting, designing, and delivering assessment tasks that reflect the characteristics of the TLU domain. The assumptions specify that test tasks and assessment conditions should mirror the TLU domain and should sufficiently represent it. They also emphasize that professional knowledge should be included in the test "to the extent that it is required by the policy environment and test purpose" (U. Knoch/ S. Macqueen 2020: 143).

Although U. Knoch / S. Macqueen (2020) only discuss validity in Chapter 5 of their six-chapter book, the issues of validity and validation should be considered from the initial stages of test design and not only after the test is in use (G. Fulcher/ F. Davidson 2007). Thus, when designing the instruments to be used in this proposed study, validation issues should be taken into consideration. The domain description assumptions of U. Knoch/ S. Macqueen's (2020) validation framework should inform the development of the questions for the interview and questionnaires (as proposed in the provisional questions included in the appendices). Thus, the evidence to be collected during the execution of this study may help to support the domain description assumptions.

2.2 Methods

The design of the proposed project consists of a multistep mixed-methods study in which qualitative data is dominant. Step 1 aims to gather useful information about the available resources and expectations of the organization developing the test or to which the test is being developed. As J. McDonough (1984) and J. Swales (1988) suggested, this needs to be taken into consideration at the beginning of the needs analysis because possible constraints may be identified from the initial stages of test development. The goal of Step 2 is to investigate the relevant skills, knowledges and processes that should be assessed in a test that aims to assess pilots' listening in isolation as stated in the ICAO policy and as discussed in the relevant literature, as well as to investigate the TLU tasks and their characteristics. The results of the data analyses of Step 2 should inform the instruments to be used in Step 3, which aims to investigate the skills, knowledges and processes assessed by a recognized test of pilots or air traffic controllers. Step 3 also aims to investigate the TLU domain tasks that are represented in the recognized test. The purpose of Step 4 is to investigate the perceptions of key stakeholders (pilots, air traffic controllers, raters, and/or researchers) on the listening construct of the test. In A. Garcia/ J. Fox (2020), pilots responded to a questionnaire on the listening construct of a test to assess pilots' listening comprehension (Phase A), and then aeronautical English experts, including raters, pilots, air traffic controllers and researchers, were interviewed (Phase B). After Step 4, you should have enough information to produce the draft design statement. Then, in Step 5, aviation English researchers or another group of experts should evaluate the design statement so that it can be improved.

Table 5 shows the types of data and their sources for each STDQ. The data gathered in Step 2 can be combined in a grid and these data should inform Step 3, and then the data collected in Step 3 can feed the grid to inform Step 4. The grid with the results of Steps 2, 3 and 4 should be used to elaborate the design statement. Data collected and analysed in Step 5 should be used to refine the design statement.

STDQ	Type of data	Sources of data or data collection methods
1	QUAL	Interview with test developer or institution requesting the test to be developed
2	QUAL	ICAO policy and ICAO test design guidelines developed by ICAEA
3	QUAL	Academic literature and research
4	QUAL	ELPAC's website, ELPAC published documents, and e-mail interview with ELPAC test developer
5	quan QUAL	Questionnaire with key stakeholders Interviews with key stakeholders
6	quan QUAL	Questionnaire with Aviation English Researchers Focus group

Table 5. Type of data and sources of data per STDQ.

A. Monteiro/ J. Fox (2022) argue that

> Test development and construct specification are strengthened when they are informed not only by theory and empirical research, but also by transdisciplinary stakeholders (e.g., pilots, ATCOs, test developers, trainers) whose expertise is rooted in varying lived experience of the construct as it plays out in actual practice. Improved construct specification leads to tests that are more aligned with the communicative needs of test takers and, as a result, have fewer unintended consequences (A. Monteiro/ J. Fox 2022: 165).

Therefore, the suggested participants in this proposed study are transdisciplinary stakeholders: pilots, ATCOs, researchers, test developers, English Language Experts (ELE) raters, with background in teaching English as a second language, Subject Matter Experts (SME) raters, who are experienced pilots or ATCOs, and aeronautical English researchers (see Table 6). Purposive sampling (L. Cohen et al. 2011) can be used in order to choose the participant to be interviewed in Step 3 (developer of a recognized test).

STDQ	Participants
1	Test developer or test requester
4	Developer of a recognized test
5	*Phase A:* airplane and helicopter pilots, ATCOs *Phase B:* Pilots, ATCOs, raters, researchers. In A. Garcia/ J. Fox (2020), 156 pilots answered a questionnaire in Phase A, and six experts were interviewed in Phase B: Four pilots (including one who holds a PhD in Linguistics and is an active researcher in the field of aeronautical English); one ATCO (all of whom have had experience working as ICAO test raters); one expert in aeronautical English who has worked as a test developer, rater trainer and test administrator and who was an active researcher in the field of aeronautical English
6	Aeronautical English researchers

Table 6. Participants in the study per STDQ.

In Step 1, qualitative data should be collected through an interview with the test developer or test requester about the resources for the development of a listening test for pilots, about the possible conditions of test administration, and about the test developers' needs. The interview can be conducted through email, as suggested by E. Dahlin (2021). The advantage of conducting the interview through email is that the participant has time to elaborate on the questions and give richer responses. Table 7 shows examples of questions that may be asked in this interview.

Since the ICAO test design guidelines developed by ICAEA state that "test instruments need to contain tasks dedicated to assessing listening comprehension, separate from tasks designed to assess speaking performance", please answer the following questions:	
1	What are the available resources for the development and administration of a test to assess pilots' listening comprehension in isolation at your organization?
2	What do you think are the possible conditions for administering a test to assess pilots' listening ability? Would it be possible for the test to be computer-based? Would there be resources available for that?
3	In your opinion, what would be the advantages and disadvantages of having a computer-based assessment?
4	What would be the advantages and disadvantages of having a separate listening test administered by the speaking test interlocutor?

Table 7. Examples of questions that may be asked in the interview with the test developer/requester.

Data gathered during this step may help to inform sections 1 and 2 of the design statement. Section 1 ("purpose of the test") should also be based on the ICAO policy and ICAO test design guidelines developed by ICAEA. Studies asking demographic questions similar to A. Garcia (2017) can also be used to inform Section 2 ("description of the test takers"). Figure 3 shows the diagram of Step 1.

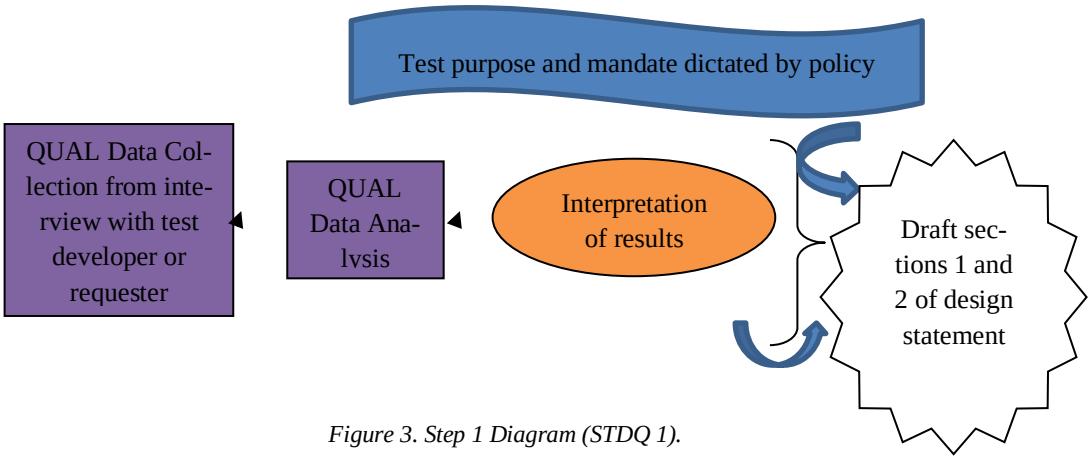

Figure 3. Step 1 Diagram (STDQ 1).

In Step 2, as seen in Figure 4, qualitative data from the policy and from the literature should be combined and used to inform both the instruments to be used in the next steps and the design statement.

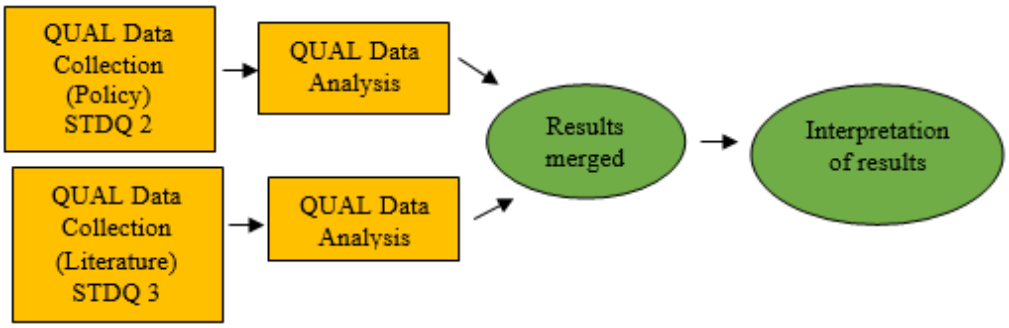

Figure 4. Step 2 Diagram (STDQ 2 and 3).

In the literature review, I suggest looking into the following topics: language assessment in general, assessment of listening, some relevant social theories, LSP, and aeronautical English, having a closer look at other studies on the construct of aeronautical English tests. Table 8 shows examples of relevant literature that could be used in this review:

Topic	Examples of relevant literature
Language assessment in general	G. Fulcher/ F. Davidson 2007, D. Douglas 2010, B. O'Sullivan 2011
Assessment of listening	G. Buck 2001, M. Rost 2016, R. Green 2017, G. Ockey/ E. Wagner 2018, J. Field 2019, O. Rossi/ T. Brunfaut 2021
Social theories	J. Lave/ E. Wenger 1991, E. Hutchins 1995a, 1995b; E. Wenger 1998, J. Jenkins 2000
LSP	D. Douglas 2000, U. Knoch/ S. Macqueen 2020
Aeronautical English	R. Yan 2009, D. Estival et al. 2016, A. Borowska 2017; J. Trippe 2018, E. Friginal et al. 2020
Research the construct of aeronautical English tests	H. Kim 2013, A. Monteiro 2019, M. Park 2021, Silva 2022

Table 8. Literature review to be conducted in Step 2 (STDQ 3).

Figure 5 shows a diagram that might be followed in Step 3. In this step, I suggest starting by collecting data about the construct of the recognized test from information available on the test's website and publicly available documents. Then, the data should be analysed. A table with the information of the test's stated construct can be produced, and the draft interview questions can be refined. Next, the e-mail interview with the test developer can be conducted. After that, the data should be analysed and merged with the results of steps 1 and 2.

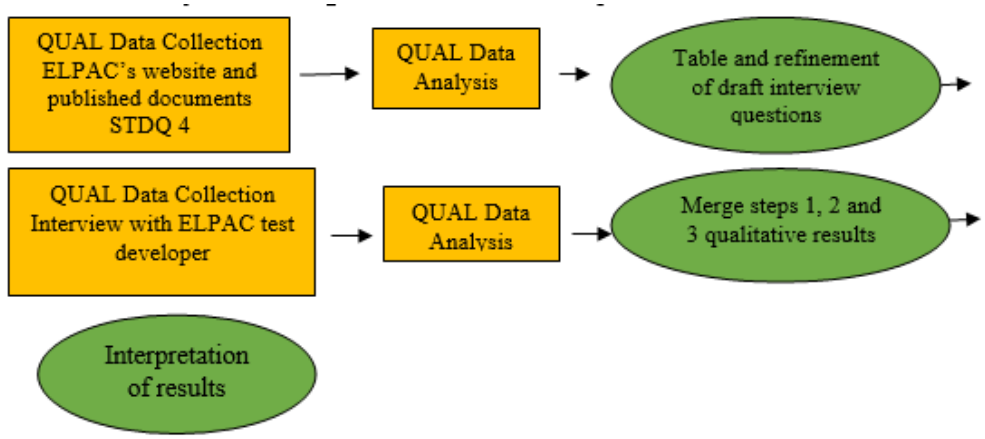

Figure 5. Step 3 Diagram (STDQ 4).

Table 9 shows examples of questions that may be asked in the interview with the developer of a recognized test. Questions 2, 5, 6, and 7 are related to the domain description assumptions 1, 2, 3, and 5, respectively, from the validity framework proposed by U. Knoch/ S. Macqueen (2020).

Please answer the following questions in as much detail as you can:	
1	What skills, knowledges, processes, and strategies does the listening test assess?
2	Do you feel assessment tasks mirror those in the TLU domain? How?
3	What do you feel the test is assessing well? And what do you feel the test is not assessing well?
4	Do you feel there is significant construct irrelevance variance (something that is being assessed but should not be)?
5	Do you feel the test is not assessing something that it should be assessing (construct underrepresentation)? In other words, do you feel that the chosen assessment tasks sufficiently represent the TLU domain?
6	Do you feel that the assessment tasks elicit and are sufficiently representative of the types of skills, knowledges and processes needed in the TLU domain?
7	How does the test incorporate the domain-specific professional knowledge? To what extent is technical knowledge included in the test?

Table 9. Examples of questions that may be asked in the interview with the recognized test developer.

Figure 6 gives an example of a diagram that might be followed in Step 4. This diagram of an empirical mixed-method study was followed in a pilot study I conducted during my PhD program. (A. Garcia/ J. Fox 2020). This study applied a two-phase explanatory sequential design, as defined by J. Creswell (2015). The first phase was the collection and analysis of the quantitative data (questionnaire answered by 156 pilots), followed by the collection and analysis of the qualitative data (interviews with six key stakeholders). The qualitative data explains or expands on the results of the quantitative data. In this mentioned study, the questionnaire was hosted on Qualtrics, whereas the interviews were semi-structured and conducted through Skype. After Step 4, the results from steps 2, 3, and 4 can be combined to draft section 3 of the design statement: the construct definition of the test.

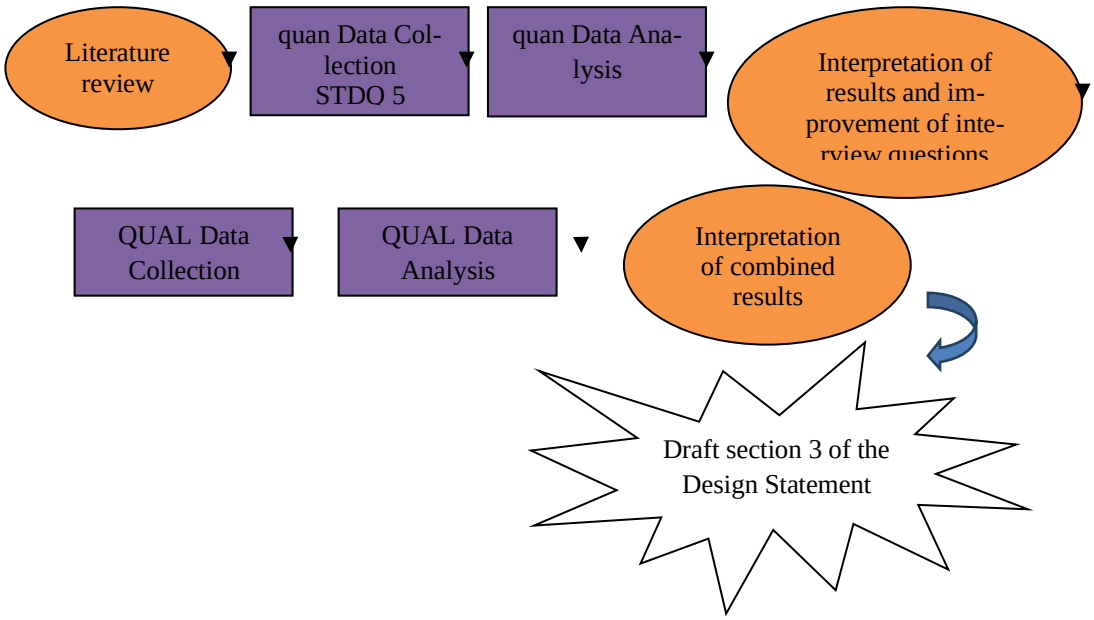

Figure 6. The Explanatory Sequential Design followed in the pilot study.

Step 5 may start with the presentation of the research project to the participants (either online or in person). In this presentation, the draft design statement detailing the test's construct definition should be explained. Next, participants can answer a questionnaire about the draft design statement. Examples of questions that can be asked in this questionnaire can be seen in Table 10. Questions 6, 7, and 8 are related to the domain description assumptions 1, 3, and 5, respectively, from the validity framework proposed by U. Knoch/ S. Macqueen (2020).

To what extent do you agree or disagree with the following statements?					
( ) I strongly disagree with it	*( )* I disagree with it	*( )* I somewhat disagree with it	*( )* I somewhat agree with it	*( )* I agree with it	*( )* I strongly agree with it.
1	*The proposed definition of the construct of a test to assess pilots' listening in isolation is complete and is not excluding any relevant skills, abilities and processes.*				
2	*The proposed definition of the construct of a test to assess pilots' listening in isolation is not including the assessment of skills, abilities and processes that are irrelevant to the TLU domain.*				
3	*The list of TLU domain tasks selected as a basis for developing assessment tasks sufficiently represent the TLU domain.*				
4	*The list of TLU domain tasks selected as a basis for developing assessment tasks are sufficiently representative of the types of skills, knowledges and processes needed in the TLU domain.*				
5	*A test to be developed on the basis of the proposed construct definition will likely assess competencies which are important for pilots when they listen to air traffic control communications.*				
6	*The assessment tasks to be created based on the proposed construction definition will likely mirror those in real-life pilot/ATCO communications.*				
7	*The list of tasks is sufficiently representative of the types of skills, knowledge and processes that pilots need when listening to radiotelephony communications.*				

Table 10. Examples of questions that may be asked in the questionnaire.

The questionnaire data should be analysed, and the focus group interview questions refined. Then, the participants should discuss in a focus group the results of the questionnaire and the improvements to the design statement. Table 11 shows examples of questions that may be asked in the focus group.

We are going to discuss the following questions:	
1	What is your opinion about the construct represented in the proposed document?
2	In your opinion, how can the proposed construct definition be improved?
3	In your opinion, does the *Design Statement* include the need to assess anything that you believe to be UNNECESSARY?
4	In your opinion, does the *Design Statement* NOT include the need to assess something that you believe to be NECESSARY?
5	Have you identified any technical or language mistake in the *Design Statement*?
6	How do you think the implementation of the test will impact aviation safety?

Table 11. Examples of questions that may be asked in the focus group.

The focus group interview should be recorded, and the recordings should be transcribed and analysed. With the results of Step 5, you can refine the draft design statement and finalize the test construct definition. Figure 7 shows the mixed-methods diagram for Step 5.

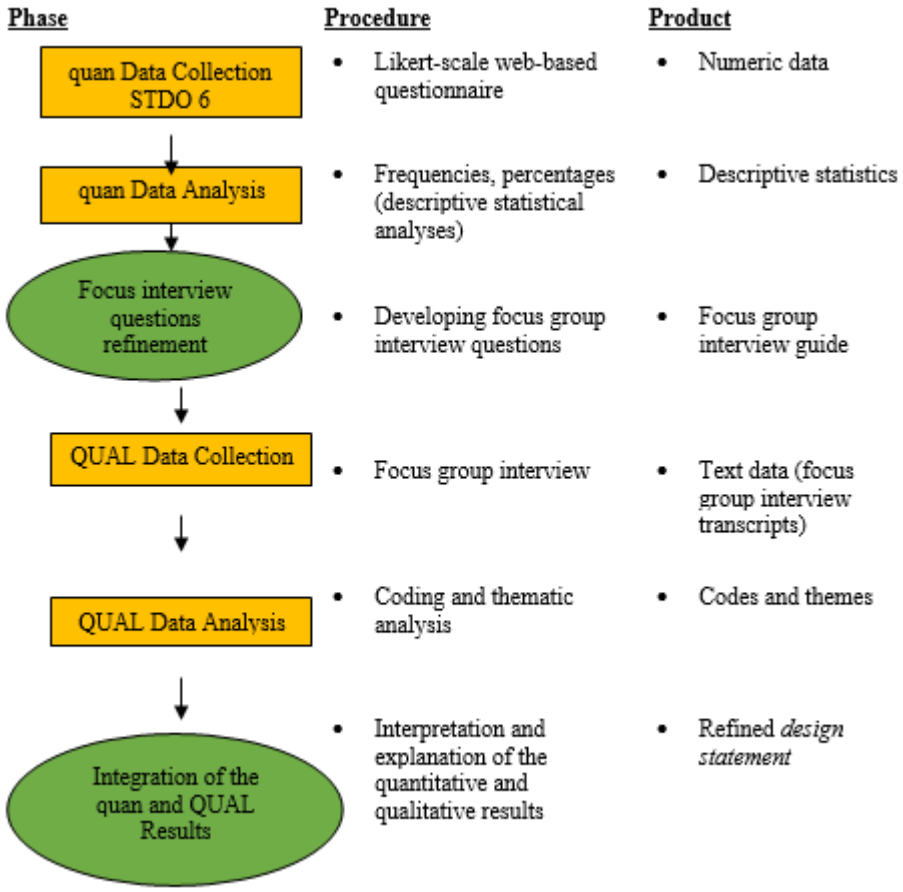

Figure 7. An Explanatory Sequential Design to be followed in Step 5 (adapted from J. Creswell/ V. Clark 2018: 85).

The qualitative data can be analysed through descriptive coding (J. Saldaña 2013) and the quantitative data generated through questionnaires can be analysed through descriptive statistical analyses. As previously mentioned, U. Knoch/ S. Macqueen's (2020) validation framework informed the development of the suggested provisional questions for the interview, questionnaire, and focus group.

3. Next steps and conclusion

U. Knoch/ S. Macqueen (2020) suggest that after the table of domain tasks is ready, test developers should examine in detail the list of tasks and check whether they are useable, suitable, workable and how likely they will result in positive test preparation behaviours and effects. In order to scrutinize this information, some questions that test developers might ask are: How much language does this task require? How easy is it to accomplish this task? Does this task pose a threat to safety? Does this task need to be adapted? Does this task rely too much on background knowledge? Is this task too specific for use in a test? How may this task influence teaching and learning? Then, test developers would be

able to start to produce the blueprint, which should include the specifications for the test and for the tasks. The development of such document is "a central and crucial part of the test construction and evaluation process" (J. Alderson et al. 1995: 9). Other critical questions to be asked that will inform the test specification and blueprint document are related to the test procedures, including the scoring method. Test developers will also need to consider how many times the test takers will be able to listen to the test prompts and how that will work (if they are to be allowed to listen to a prompt more than once, or if they will have to show they need to listen to it again by asking for clarification, or if the recordings will be played twice anyway), as in real-life situations pilots are expected to ask for clarification. Also, test developers will need to decide on what accents to include. In relation to this, E. Wagner (2022) points out:

> The issue of accent variety is an important consideration for L2 listening test developers because they need to consider what accent or variety to use on their tests. The obvious answer would seem to be to use the "standard" variety of the language where the test is taking place, but in an age of globalization and multiculturalism, choosing the most appropriate accent varieties to use on an L2 listening test can be challenging (E. Wagner 2022: 228).

In designing a test to assess Brazilian pilots, test developers could check, for example, the most frequent international routes that Brazilian pilots take in order to define the most frequent accents to be included in the test and its proportion. Once the draft of test specifications and blueprint document is written, test developers can create a sample test and write enough items to pilot the test with a representative number of test takers. Aeronautical English corpora can be very useful in the creation of test items (A. Pacheco et al. 2020).

After having given suggestions of future research, I need to acknowledge some limitations of the project explained in this paper. As most parts of the gathered data are qualitative, the results might be biased by the subjective interpretation of the researcher. Also, there are several other ways a needs analysis could be conducted. This is just one possible way of doing it. Furthermore, I recognize that it is not possible to fully address all research questions or to review all existing relevant literature. This reminds me of U. Knoch/ S. Macqueen's (2020) reflection:

> There is never a definite outcome from a needs analysis, rather researchers are required to decide, within a policy environment and certain fixed test requirements, how workplace communication can best be presented in a series of test tasks. There is therefore no point in searching for the 'truth' in the data collected, but, … , the researcher needs to create their current-best-shot at what a test should look like (U. Knoch/ S. Macqueen 2020: 108).

To conclude, needs analysis should not only happen in the initial stage of test development. It should be repeated to make sure the test continues to represent the TLU adequately (U. Knoch/ S. Macqueen 2020). Moreover, in spite of the different goals practitioners and language test researchers may have, a clear definition of the construct is extremely important, as L. Bachman (2007) thoughtfully points out:

> Perhaps the most important distinction between the roles of language testing researcher and practitioner is that of purpose, or goal. The language testing researcher's

> goal, I believe, is to better understand, inter alia, the psychological and contextual factors that affect performance on language assessments, the types of language use that language assessments elicit, the relationship between language use elicited in assessments and that created in real-life settings, and the relationship between the abilities engaged in language assessments and those engaged in real-life settings. I would argue that the goal of the language testing practitioner, on the other hand, is to design and develop language assessments that are useful for their intended purposes. In either role, I believe that it is essential that we clearly define what it is we want to measure or what we want to investigate (L. Bachman 2007: 66).

If, like me, you are both a practitioner and a language test researcher, I hope that you can reconcile both of our goals by conducting an academically sound study that aims to clearly define what a listening test for pilots should assess in order to inform your practice as test developers. R. Green (2017: 29) states that "defining the construct accurately and reliably is arguably one of the most important responsibilities of test designers". I trust we can accomplish this goal responsibly.

References

Alderson, J.C./ C. Clapham/ D. Wall (1995), *Language test construction and evaluation.* Cambridge.

Bachman, L.F. (2007), *What is the construct? The dialectic of abilities and contexts in defining constructs in language assessment*, (in:) J. Fox et al. (eds.), "Language Testing Reconsidered", Ottawa, 41–71.

Bachman, L.F. (2015), *Justifying the use of language assessments: Linking test performance with consequences*, (in:) "JLTA Journal" 18, 3–22.

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

Bachman, L.F./ A.S. Palmer (2010), *Language assessment in practice: Developing language assessments and justifying their use in the real world.* Oxford.

Borowska, A.P. (2017), *Avialinguistics: The Study of Language for Aviation Purposes.* Peter Lang: Frankfurt am Main.

Brown, J.D./ T. Rogers (2002), *Doing second language research.* Oxford.

Buck, G. (2001), *Assessing Listening.* Cambridge.

Chalhoub-Deville, M. (2003), *Second language interaction: current perspectives and future trends,* (in:) "Language Testing" 20(4), 369–383.

Chapelle. C. (1998), *Construct definition and validity inquiry in SLA research*, (in:) L.F. Bachman / A.D. Cohen (eds.*)*, "Interfaces between second language acquisition and language testing research", Cambridge, 32–70.

Chapelle, C.A./ M.K. Enright/ J.M. Jamielson (2008), *Building a validity argument for the test of English as a foreign language.* New York.

Chapelle, C./ H. Lee (2021), *Understanding argument-based validity in language testing*, (in:) C. Chapelle/ E. Voss (eds.), "Validity argument in language testing: Case studies of validation research", Cambridge, 19–44.

Cohen, L./ L. Manion/ K. Morrison (2007), *Research methods in education* (6th Ed.). New York.

Creswell, J.W. (2015), *A Concise introduction to mixed methods research*. Thousand Oaks.

Creswell, J.W./ V.L.P. Clark (2018), *Designing and conducting mixed methods research* (3rd ed.). Thousand Oaks.

Dahlin, E. (2021), *Email interviews: A guide to research design and implementation*, (in:) "International Journal of Qualitative Methods" 20, 1–10.

Dörnyei, Z. (2007), *Research methods in applied linguistics: quantitative, qualitative and mixed methodologies*. Oxford.

Douglas, D. (2000), *Assessing languages for specific purposes*. Cambridge.

Douglas, D. (2001), *Three problems in testing language for specific purposes: Authenticity, specificity and inseparability*, (in:) C. Elder et al. (eds), "Experimenting with uncertainty: Essays in honour of Alan Davies", 45–52. Cambridge.

Douglas, D. (2010), *Understanding language testing*. London.

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

Field, J. (2019), *Rethinking the second language listening test – from theory to practice.* Sheffield.

Friginal, E./ E. Mathews/ J. Roberts (2020), *English in global aviation: Context, research and pedagogy*. London.

Fulcher, G./ F. Davidson (2007), *Language testing and assessment an advanced resource book*. London.

Garcia, A.C.M. (2017), *The end-users' perceptions of test content: Is it working?* Workshop delivered at the International Civil Aviation English Association International Workshop, Dubrovnik.

Garcia, A.C.M. / J. Fox (2020), *Contexts and constructs: Implications for the testing of listening in pilots' communication with air traffic controllers,* (in:) "The ESPecialist" 41(4), 1–33.

Green, R. (2017), *Designing listening tests: A practical approach*. London.

Harding, L. (2015), *Book review: Examining listening: research and practice in Assessing Second Language Listening*, (in:) "Language Testing" 32(1), 121–124.

Hutchins, E. (1995a), *Cognition in the Wild*. Cambridge.

Hutchins, E. (1995b), *How a cockpit remembers its speeds*, (in:) "Cognitive Science" 19, 265–288.

ICAEA (n.d.a), *ICAO LPR Test Design Guidelines – Criterion 3.* Accessed on 24 June 2022 from https://www.icaea.aero/projects/icao-lpr-tdg/guidelines/tdg-criterion-3/.

ICAEA (n.d.b), *ICAO LPR Test Design Guidelines – Criterion 2.* Accessed on 20 December 2022 from https://www.icaea.aero/projects/icao-lpr-tdg/guidelines/tdg-criterion-2/.

ICAEA (n.d.c), ICAO Test Design guidelines and explanations to evaluate the design of ICAO LPR Tests. [Accessed on 25.07.2023 from https://www.icaea.aero/projects/icao-lpr-tdg/guidelines/].

ICAO (2010), *Manual on the implementation of ICAO language proficiency requirements (Doc 9835)* (2nd ed.). International Civil Aviation Organization.

ICAO (2020), *Annex 1 to the Convention on International Civil Aviation – Personnel licensing* (13th ed.). International Civil Aviation Organization.

Jacoby, S./ T. McNamara (1999), *Locating competence,* (in:) "English for Specific Purposes" 18(3), 213–241.

Jenkins, J. (2000), *The phonology of English as an international language*. Oxford.

Kane, M. (2002), *Validating high-stakes testing programs,* (in:) "Educational Measurement, Issues and Practice" 21(1), 31–41.

Kane, M. (2006), *Validation,* (in:) R. Brennan (ed.), Educational measurement (4th ed.), 17–64. Westport.

Kim, H. (2013), *Exploring the construct of radiotelephony communication: A critique of the ICAO English testing policy from the perspective of Korean aviation experts,* (in:) "Papers in Language Testing and Assessment" 2(2), 103–110.

Knoch, U./ C. Chapelle (2018), *Validation of rating processes within an argument-based framework*, (in:) "Language Testing" 35(4), 477–499.

Knoch, U./ S. Macqueen (2020), *Assessing English for professional purposes*. Milton.

Lave, J./ E. Wenger (1991), *Situated Learning: Legitimate Peripheral Participation*. Cambridge.

McDonough, J. (1984), *ESP in perspective: a practical guide*. London.

Messick, S. (1989), *Validity,* (in:) R.L.Linn (ed.), Educational measurement (3rd ed.), 13–103. Phoenix.

Milanovic, M./ C. Weir (2013), *Series Editors' note.* (in:) A. Geranpayeh/ L. Taylor (eds.), "Examining listening: Research and practice in assessing second language listening", ix-xvi. Cambridge.

Moder, C.L/ G.B. Halleck (2021), *Designing language tests for specific purposes*, (in:) G. Fulcher/ L. Harding (eds.), "The Routledge Handbook of Language Testing". New York.

Monteiro, A.L.T. (2019*), Reconsidering the measurement of proficiency in pilot and air traffic controller radiotelephony communication: from construct definition to task design.* (Unpublished PhD's dissertation). Carleton University.

Monteiro, A.L.T. / J. Fox (2022), *Clarifying the testing of aural/oral proficiency in an aviation workplace context: social theories and transdisciplinary partnerships in test development and validation,* (in:) J. Fox/ N. Artemeva, "Reconsidering context in language assessment: Transdisciplinary Perspectives, Social Theories, and Validity", 163–197. Milton.

Ockey, G./ E. Wagner (2018), *Assessing L2 listening: moving towards authenticity*. Amsterdam.

O'Sullivan, B. (2011), *Language testing: theories and practices*. Basingstoke.

Pacheco, A./ A.L.T. Monteiro/ A.C.M. Garcia/ M. Prado/ P. Tosqui-Lucks (2020, November), *Using real aviation communications to create tasks for training and testing* [Webinar]. International Civil Aviation English Association. https://www.icaea.aero/webinars/webinars-2020/.

Park, M. (2021), *Domain definition inference for a virtual interactive aviation English test (VIAET) for military air traffic controllers,* (in:) C. Chapelle/ E. Voss (eds.), "Validity argument in language testing: Case studies of validation research", 73–95. Cambridge.

Rossi, O./ T. Brunfaut (2021), *Text authenticity in listening assessment: Can item writers be trained to produce authentic-sounding texts?* (in:) "Language Assessment Quarterly", 1–21.

Rost, M. (2016), *Teaching and researching listening* (3rd ed.). London.

Saldaña, J. (2016), *The coding manual for qualitative researchers* (3rd ed.). Los Angeles.

Silva, A.L.B.C. (2022*), Avaliação de proficiência em inglês para pilotos da Esquadrilha da Fumaça: Da análise de necessidades ao desenho de um exame*. [*English language proficiency assessment for pilots from Esquadrilha da Fumaça: From needs analysis to test design]* (Published PhD's dissertation). Universidade Estadual de Campinas. [Accessed on 05.07.2022 from https://repositorio.unicamp.br/Acervo/Detalhe/1244211?returnUrl=%2FIndicador%2FIndex%3Fc%3D1244211].

Swales, J. (1988), *Episodes in ESP*. New York.

Tosqui- Lucks, P./ A.L.B.C. Silva (2020), *Da elaboração de um glossário colaborativo à discussão sobre os termos inglês para aviação e inglês aeronáutico*. [From writing a collaborative glossary to discussing the terms "aviation English" and "aeronautical English"], (in:) "Estudos Linguísticos" 48(4), 97–116.

Trippe, J.E. (2018), *Aviation English is distinct from conversational English: Evidence from prosodic analyses and listening performance*. (PhD's dissertation). University of Oregon.

Wagner, E. (2022), *Assessing listening*, (in:) G. Fulcher/ L. Harding (eds.), "The Routledge Handbook of Language Testing". New York.

Weir, C.J. (2005), *Language testing and validation: An evidence-based approach*. London.

Wenger, E. (1998), *Communities of practice: Learning, meaning, and identity*. Cambridge.

Yan, R. (2009), *Assessing English language proficiency in international aviation: Issues of reliability, validity and aviation safety*. PhD's dissertation. University of Louisiana at Lafayette.