Indonesian Journal of EFL and Linguistics

eISSN: 2503-4197, pISSN: 2527-5070
Available online at:
www. indonesian-efl-journal.org
http://dx.doi.org/10.21462/ijefl.v11i1.1015

Designing A Speaking Test for Special Purposes (SP) in Hiring Experts for a Company

Md. Abdus Salam
Jagannath University, Bangladesh
e-mail: salam@eng.jnu.ac.bd

Abstract:

Designing speaking tests for special purposes involves contentious debates regarding specific contexts, company goals, and learner requirements, and hence, requires further research. This article designs a speaking test for the specific purpose of recruiting company professionals. At present, multinational companies around the globe are evolving day by day, and their sister concerns are getting diverse perspectives. The research objective of this article is to design a speaking test. This study requires finding an answer to the research question: how to design a valid and reliable speaking test for recruiting company professionals? This study applied a qualitative data analysis approach to find an answer to the research question, in which research articles, books, periodicals, and magazines were treated as secondary, and the speaking test design was primary data. This study finds that multiple challenges are present in designing a speaking test, keeping in mind the reliability and validity of the tests, like test-time, assessment-criteria, validity, and reliability. The findings of this article will be helpful and have positive impacts for the examinees as well as the employers. Observing the necessities, this study recommends further research for designing a speaking test, keeping real-world relevance in all contexts.

Keywords: *Hi-stakes test, reliability, speaking-test-design, validity*

1. INTRODUCTION

Tests such as the Testing English for Special Purposes (TESP), International English Language Testing System (IELTS), Situational Judgment Tests (SJTs), Test of English as a Foreign Language (TOEFL), English Speaking for Other Level (ESOL), Graduate Record Examinations (GRE), General Proficiency Test (GPT), Occupational English Test (OET), etc. are used to

assess language proficiency in macro skills, which include speaking, listening, reading, and writing. Each test has a unique methodology, relevance, implications, and test specification. The study discussed in this article provides commentary on the nature of speaking proficiency, or speaking for Special Purposes (SP) test, in the context of employing specialists for a corporation. As Jung Youn (2023) found importance in validity in terms of contexts for the speaking test design.

This article's initial section discusses the rationale behind selecting texts and exercises that are validated by pertinent theories. An examination of the difficulties this article encountered when creating the test comes next. The pertinent literature on language testing and evaluation will subsequently be examined and discussed. The article's changes will then be evaluated in the context of the literature review. After the test design has been successfully completed, the article will conclude with a discussion of the recommendations.

The SP speaking test was chosen and designed in this article for a number of reasons, which will be examined in the following. For a number of validity-related factors, this article has selected the SP speaking exam to hire experts. The question of whether speaking competence is more crucial for accurately assessing language competency is one of many reasons, and it is contentiously debatable. First, this study created an SP speaking test based on the relationship between detached features of the spoken language produced by the test takers and holistic scores given by raters in order to hire experts and service providers for a firm (Iwashita et al. 2008). Test takers' performance in ten different tasks and ten different proficiency levels is rated or marked using a variety of metrics, including fluency, coherence and organisation, clarity or volume of speaking, vocabulary, grammar and structure, pronunciation, eye contact, body language, usage of phrases and linkers, task achievement or completion. These metrics have a significant impact on a company's top-performing employees. Second, a crucial component of applied linguistics is second language (L2) proficiency. However, there is much disagreement over the necessary level of competency in the pertinent context. In general, the Occupational English Test (OET) will not benefit from general proficiency in everyday English language usage. Healthcare workers' language and communication skills are evaluated by OET (Davidson, 2018). OET speaking subtests are tailored to each individual and are intended to evaluate proficiency in using English in a pertinent professional setting. Travelling and daily communication are two situations where general proficiency is beneficial. Conversely, distinct terminology and proficiency are needed for Academic English, Business English, Aviation English, Military English, Life-guard English, and English for Tourism. Because of this, the American Council on the Teaching of Foreign Languages (ACTFL) creates competence tests that are utilized in American academia at the college and university levels (Breiner-Sanders, et al., 2000). According to Lumley (1998), "OET uses' speaking 'materials specially designed for' the test-taker's 'profession." Tests like the TOEFL, GRE, and IELTS are taken in order to pursue higher education in English-speaking nations. Globally, multinational corporations are searching for a self-assured worker who can market the company's goods or services. Therefore, a variety of multinational corporations, including Adidas, Asus, Amazon, Bata, Barclays, BRAC, BMW, Coca-Cola, Chevron, Dell, DHL, Epson, FedEx, Facebook, Google, HSBC, Honda, KFC, IBM, Nike, LG, Nokia, Sony, Samsung, Standard Chartered, Square, Starbucks, Tesco, Tata, Toyota, Toshiba, Unilever, Uber, United Airlines, Yahoo, Whirlpool, and others, require workers with varying degrees of language proficiency. For large international corporations, writing, reading, and listening assessments are insufficient to find qualified professionals.

Shin (2022) suggests an alternative speaking test design with fewer items. It means that speaking tests with fewer items sometimes puts forward the validity, but in case of hiring company professionals speaking test requires more validity and reliability on which this article is framed. However, Tests involving writing, reading, and listening require a lot of time and money. For the reasons mentioned above, this article has selected the SP speaking test, which may be used by all multinational corporations worldwide to find the most qualified candidates for their workforce. Thirdly, it is impossible to accurately assess and judge test takers' confidence, demeanor, and comprehension level using written, reading, and listening assessments. However, the test takers' body language and eye contact make it easy to determine their level of comprehension and confidence. According to analysis, an examinee cannot persuade consumers or customers with a respectable social context if they lack appropriate eye contact and body language; if they answer at a low volume, they will not be audible to clients in pertinent context; and if they occasionally request repetition or reformation of the questions, it suggests that they will not comprehend the client's speech in a real-world work setting. According to MacIntyre et al. (1998), confidence, social attitude, and comprehension levels appear to be similar in all L1 and L2 contexts.

It goes without saying that these levels of comprehension are essential for marketing a business's goods or services. This SP speaking exam can be used in any setting. In order to identify different linguistic levels of proficiency based just on the candidate's speaking skills, this test was chosen for the in-depth analysis of test-taker speech in this article. This article tries to find an answer to the research question: "How to design a valid and reliable speaking test for recruiting company professionals?"

2. LITERATURE REVIEW

Language proficiency is linked to the user's understanding of the structures and characteristics of the communication context, according to Bachman (1990). Interviews are a typical way to assess speaking abilities. Hatipoğlu (2021) found "speaking is a productive skill" (p. 124), and for this reason, speaking tests should be designed empirically. It reads that to judge a productive language skill like speaking, validity, reliability, and context are very crucial and important.

Hughes (2003) highlights a potentially significant disadvantage:

> "The candidate typically speaks to the tester as a superior and is reluctant to take the initiative because of their connection."

This article has created an interactive test in the test specification to prevent this kind of problem. "Yes/No" questions should be avoided in all test parts, according to the test format. The question will look something like this:

1. Could you briefly introduce yourself?
2. Could you tell us how and why?
3. Could you share your thoughts with us?

Interviewers may enquire about further details:

1. What do you mean precisely?
2. Could you elaborate a bit more on that?
3. What would be a suitable example of that?
4. Give us further details.

Interviewers may suddenly shift the subject to observe how the candidate responds. Interviewers may seem perplexed and pose questions such as these:

1. I apologize, but I don't really understand you.

The candidate may be invited by the interviewer to ask a question:

1. Do you have any questions to ask us? Hughes (2003)

The biggest problem for this article was designing the speaking test without any drawbacks in order to preserve the test's face validity. Khabbazbashi et al. (2022) indicate "promoting speaking practice in classrooms" (p. 141) and find challenges in assessing young learners by applying technological innovation. This article is to design face to face speaking test for the examinees.

Test takers may be asked questions in a variety of ways during speaking and interview formats. Candidates may be asked to take on a role in a scenario, interpret a particular sample, give a prepared monologue, read aloud a sample, engage in discussion and interaction with other candidates, respond to audio or video recordings, describe a scenario, or participate in a simulated conversation (Hughes, 2003). During role-playing, there may be a quick item:

a) You wish to go by plane from London to Paris on March 13 and back a week later. Obtain all the information required to select your flights from the travel agent.
b) You want your pocket money to be increased by your mother (tester). She rejects the notion. Make an effort to persuade her to reconsider.
c) On a night when you would prefer to stay home and watch the final episode of a television serial, a friend asks you to a party. Thank the friend (the tester) and politely decline (Hughes, 2003).

However, in real-world experience, role-playing could cause the candidate to lose focus and stray from the subject. When interpreting:

> "The native speaker wishes to extend an invitation to a foreign guest to join them for dinner. In addition to serving as an interpreter for the ensuing conversation, the candidate must deliver the invitation" (Hughes, 2003).

When the applicant tries to explain what the visitor is saying, their comprehension may be evaluated. However, it is challenging to get enough data on candidates' comprehension and production skills. According to Hughes (2003), prepared monologues are "frequently misused." As an alternative, this could be useful for introducing a subject that the applicants would eventually need in their real-world lives. Reading aloud is a useful method for evaluating test takers' tone and pronunciation. However, the L1 and L2 speakers' readings would differ from one another. Additionally, under this test specification criterion, speaking ability will be hampered by reading proficiency. Ginting et al. (2023) criticise the "reliability and validity" (p. 138) of the IELTS speaking test. In the case of the IELTS test, the speaking test is assessed on a different date and at a different place from the IELTS reading and writing tests. Although the IELTS speaking test is accurate. Ginting et al. (2023) mention:

> "Subjective factors like candidate preferences and marker performance will have a significant impact on the entire testing process and evaluation findings." (Ginting et al. 138)

Most of the time, candidates are afraid since they believe they are dealing with seniors. In this sense, a candidate's ability to use natural language may also be evaluated through interactions with other candidates. They will believe that they are communicating with equals while

interacting with other applicants, and the test specification validates this assumption. However, one candidate's performance is likely to be influenced by the others. Another type of speaking skill proficiency test is a computer-generated speaking test. Candidates respond to the same computer-generated or audio-/video-recorded information by speaking into a microphone. This format is frequently referred to as "semi-direct." There is no way to disrupt the candidate during the speaking portion of this kind of exam. Therefore, the semi-direct computer-generated speaking test cannot assess a candidate's attitude, demeanor, or degree of confidence. Additionally, Educational Testing Services (ETS) created the Test of Spoken English (TSE), which proposes:

i) After viewing a basic town map, candidates are asked to: (a) suggest a visit to one of the buildings and explain why; (b) provide directions to the movie theatre; and (c) summarise their favourite film and explain why.
ii) Candidates are shown a series of images of a man sitting on a freshly painted park bench. They are asked to: (a) tell the story; (b) explain how the accident could have been prevented; (c) imagine that the accident happened to them and they have to convince the dry cleaners to clean their suit the same day; and (d) list the benefits and drawbacks of using newspapers and television as news sources (the man in the pictures is reading a newspaper on the park bench!).
iii) Candidates are asked to define a key phrase in their field of study, explain the data presented on a graph, and evaluate its implications, as well as whether or not maintaining animals in zoos is desirable.
iv) Candidates receive printed trip details with some handwritten revisions. They have to describe the modifications in a presentation to the group of travellers.
v) Candidates are informed of the time allotted for studying the material and the duration of their speech (Hughes, 2003).

Anyone interested in creating tape-meditated speaking assessments can benefit from the models mentioned above, which are suggested by ARELS and TSE. Nevertheless, TSE doesn't actually try to evaluate interaction skills. After resolving these shortcomings, the second part of the test that this article has created is titled "Simulated speaking or conversation," in which the candidate is given information about the business and its goods and services. Gong (2023) indicates “positive washback” (p. 1) and speaking test assessment and finds correlations between washback and test assessment. The Association of Recognised English Language Schools (ARELS) created methods including "simulated conversation." Candidates will have three minutes to prepare before presenting or demonstrating the business and its goods or services, treating the interviewers as clients and customers and attempting to close deals. In light of all of this, this article has selected "make the sale" and "meeting and greeting" as the near-total test specifications for the SP speaking exam. Interviewers will not intervene to assess a candidate's degree of confidence during any test procedure, especially the speaking interview mentioned above. However, in the exam structure this article has created, the interviewers will take on the role of a client, customer, or consumer, interjecting during the candidate's response and changing the subject abruptly to gauge the candidate's degree of confidence. Lastly, the test's design has relevant implications "in language testing research contexts" and offers "important insights" into the nature of speaking skill "as it develops and can be measured in" evaluating speaking skill (Iwashita et al., 2008). The following paragraph will go into more detail about the difficulties this author had when creating the SP speaking test.

Reviewing the literature, the researcher finds a research gap in designing the speaking test for hiring company professionals. Hence, in order to answer the research question, "What makes a speaking test valid for recruiting company professionals?" this paper reviews the literature mentioned above.

3. RESEARCH METHODOLOGY

This study adopts a qualitative research design grounded in document analysis and test-development inquiry to answer the central research question: How can a valid and reliable speaking test be designed for recruiting company professionals? Qualitative design is suitable because the study aims to interpret, analyze, and synthesise conceptual frameworks and test-design principles, rather than compute statistical scores. This follows established understandings of qualitative inquiry in language assessment, where test specifications, TLU domains, and rating scales can serve as analyzable qualitative data (Bachman, 1990; Hughes, 2020).

The study draws on two categories of data. The first is secondary and tertiary data, consisting of research articles, books, reports, and online academic sources related to speaking assessment, test design, CEFR-based descriptors, and English for Specific Purposes (Breiner-Sanders et al., 2000; Fulcher & Reiter, 2003). These data inform the conceptualisation of construct validity, reliability, authenticity, interactiveness, and impact (Bachman & Palmer, 1996; Hughes, 2020). The second category is primary data, which refers to the evolving drafts of the Speaking for Special Purposes (SP) test itself—its specifications, tasks, prompts, scoring criteria, and administration protocols. These components are analysed qualitatively as developing artefacts, iteratively improved across the different phases of design based on literature and expert judgement.

The procedures begin with a needs analysis derived from literature on English for Occupational Purposes and speaking proficiency requirements in multinational companies. Research on high-stakes speaking tests—such as ACTFL OPI, IELTS Speaking, OET Speaking, and institutional placement tests—provides models for defining the TLU domain and operationalising proficiency levels (Bygate, 1987; Davidson, 2018; Zhang & Head, 2010). Based on the communicative demands of professional settings (e.g., product explanation, customer persuasion, interaction management), CEFR B2 is selected as the minimum proficiency target for test-takers.

Next, test specifications are constructed to provide a blueprint of the assessment. These include test purpose, candidate profile, TLU domain, timing, task types, rating scale, scoring rules, interviewer roles, and recording procedures. To operationalise the speaking construct, the study identifies 10 observable dimensions: fluency, coherence/organisation, clarity/volume, vocabulary choice, grammatical control, pronunciation, body language, eye contact, use of linkers/signposting, and task achievement. These dimensions are aligned with established descriptors found in speaking test research, including interactional competence, discourse management, and lexical-grammatical range (Galaczi, 2008, 2013; Hughes, 2003).

To strengthen reliability, the test uses dual-rater scoring, supported by audio recording for moderation and post-hoc rating. If the score discrepancy between the two raters is 20% or more, a third rater reviews the recording to determine an adjusted final score. Raters undergo brief training using benchmark samples and guided calibration, following suggestions in reliability-focused assessment literature (Lumley, 1998; Spolsky, 1995).

Data analysis follows a thematic document-analysis approach, identifying concepts related to test usefulness (validity, reliability, authenticity, interactiveness, practicality). Test components—specifications, tasks, and rating descriptors—are compared iteratively with these themes using constant comparison techniques (Gong, 2023; Khabbazbashi et al., 2022). Alignment checks are also conducted with CEFR descriptors, existing high-stakes speaking tests, and workplace interaction literature to ensure content and construct validity.

4. RESULTS

4.1 The Format of the Test

Title: Creating a Speaking for Special Purpose (SP) Exam to Hire Professionals for a Business (Service Provider or Sales Professional).

Skill	**Speaking**
Level of the CEFR	B2
Format for Paper	There are two parts to the test. There will be an audio recording of the test.
Timing	15 minutes
The quantity of test-takers	As many as apply (one applicant at a time)
The test's objective	To assess a candidate's ability to use the English language for communication, persuasion, and product or service sales
Test for/Who the Examinee Is	Prospective employee (marketing candidate with a bachelor's degree in any field)
Domain or Target Language Use (TLU)	Retail Marketing and Sales
Administration of tests	• Individual interviews • Two interviewers, two markers, and two examiners (trained is advised) • Candidates will receive information about the company and its products.
Test sections and format	• Greetings and Meeting (5 minutes) Questions: (Avoid asking yes/no questions.) 1. Could you briefly introduce yourself? 2. Could you tell us how and why? 3. Could you share your thoughts with us? Enquiries for further details: 1. What do you mean specifically? 2. Could you elaborate on that a bit more? 3. What would be a suitable illustration of that? 4. Give us additional details. *N.B. Interviewers may suddenly shift the subject to observe how the candidate responds.* • Three minutes for preparation • Make the product or service sale (7 minutes) Interviewers could seem incomprehensible. Inquiry: 1. I apologise, but I don't really understand you. The candidate may be encouraged to ask questions to the interviewer: 1. Do you have any questions to us? *N.B.: Interviewers will pretend to be customers, clients, or consumers.*
Which construct or constructs (marking and grading) need to be evaluated?	• Fluency, Coherence and Organisation, Clarity/Volume of Speaking, Vocabulary, Grammar and Structure, Pronunciation, Eye Contact, Body Language, Phrases/Linkers/Signposting languages, Task Achievement/Completion • 50 marks in total • Pass/Selected: 60% or higher; Fail/Not Selected: less than 60%

Reliability	• The same test (good or service) on the same day-same test duration • Audio recording for rubric • Double-blind interviewers and markers (the candidate's final score will be the average of two markers' scores) • The audio recording will be sent to a different 3rd marker if the difference between the markers is 20% or more. In this instance, the candidate's mark will be the average of three distinct markers.)
Test security	• Password-protected or encrypted • Ten characters will make up the password, five of which will be provided by one marker and the remaining five by another.

2. Overall CEFR Descriptors (B2):

Spoken Interaction

Able to provide the relevant and pertinent information in straightforward, everyday situations; Can manage brief social interactions, but is rarely able to comprehend enough to carry on a conversation on their own.

Spoken Production

Can provide a brief succession of straightforward statements and sentences that are connected to the questions posed in order to provide a basic description or presentation about oneself, living or working conditions, daily routines, likes/dislikes, etc.

Part 1—Meeting and Greeting

Performance Indicator: Give a brief introduction. Discuss your personal information. Discuss your areas of strength and weakness.

Descriptors from CEFR: able to make social contact by introducing oneself, saying hello and goodbye, and expressing gratitude. Generally speaking, he or she is able to comprehend conventional, clear communication on topics they are familiar with, as long as they are able to occasionally request repetition or reformulation.

Preparation: No preparation needed.

Prompt: Questions concerning personal information (name, age, outcome, nationality, experience, etc.) are posed to interviewees. This section should not take longer than five minutes, although the number of questions varies.

Part 2—Make the Sale/Provide The Service/Convince The Customer or Consumer

Performance Indicators: Give the interviewers a coherent and fluid explanation of the company's product or service.

CEFR Descriptors: Can respond to simple follow-up enquiries if he or she is able to request repetition and if assistance in crafting a response is feasible; able to succinctly and coherently explain; able to use low-frequency words and talk in a variety of forms; able to communicate using phrases, linkers, and signposting language.

Preparation: Three minutes will be given for preparation, along with the necessary product and company information.

Interviewer's Mark Sheet

The interviewee's name and serial number	Fluency Marks: *5*	Organisation and Coherence Marks: *5*	Speaking Volume and Clarity Marks: *5*	Vocabulary Marks: *5*	Grammar and Structure Marks: *5*	Pronunciation Marks: *5*	Eye contact Marks: *5*	Body language Marks: *5*	Phrases/Linkers/Signposting languages Marks: *5*	Task Achievement/ Completion Marks: *5*	Total Marks: *50*	Comment (Selected/Not Selected)

Notes: This test may be administered anywhere in the world where the English language is the medium of communication as a second language (L2) or L1 (first language).

Assessment Criteria
Proficiency Level B2

Criteria	100%	80%	60%	40%	20%
	5	4	3	2	1
1. Fluency	a) Communicates fluently most of the time.	a) Usually communicates fluently most of the time.	a) Sometimes speaks with ease and fluency.	a) Occasionally speaks with ease and fluency.	a) Rarely speaks with ease and fluency.
2. Coherence and Organization	a) In most cases arranges speech to create a cohesive dialogue.	a) Usually arranges speech to create a cohesive dialogue.	a) Sometimes arranges speech to create a cohesive dialogue.	a) Occasionally arranges speech to create a cohesive dialogue.	a) Rarely arranges speech to create a cohesive dialogue.
3. Clarity/ Volume of Speaking	a) In most cases speaks in a clear, intelligible voice that interviewers can understand.	a) Usually speaks in a clear, intelligible voice that interviewers can understand.	a) Sometimes speaks in a clear, intelligible voice that interviewers can understand.	a) Occasionally speaks in a clear, intelligible voice that interviewers can understand.	a) Rarely speaks in a clear, intelligible voice that interviewers can understand.
4. Vocabulary	a) Most of the time appropriately employs a suitable vocabulary.	a) Usually appropriately employs a suitable vocabulary.	a) Sometimes appropriately employs a suitable vocabulary.	a) Occasionally appropriately employs a suitable vocabulary.	a) Rarely appropriately employs a suitable vocabulary.
5. Grammar and Structure	a) In most cases employs a wide range of grammatical and conversational constructs.	a) Usually employs a wide range of grammatical and conversational constructs.	a) Sometimes employs a wide range of grammatical and conversational constructs.	a) Occasionally employs a wide range of grammatical and conversational constructs.	a) Rarely employs a wide range of grammatical and conversational constructs.
6. Pronunciation	a) In most cases appropriately pronounces the words. b) Most of the time employs appropriate stress	a) Usually appropriately pronounces the words. b) Usually employs appropriate stress	a) Sometimes appropriately pronounces the words. b) Sometimes employs appropriate stress	a) Occasionally appropriately pronounces the words. b) Occasionally employs appropriate stress	a) Rarely appropriately pronounces the words. b) Rarely employs appropriate stress

	and intonation.	and intonation.	and intonation.	and intonation.	and intonation.
7. Eye contact	a) In most cases makes eye contact well.	a) Usually makes eye contact well.	a) Sometimes makes eye contact well.	a) Occasionally makes eye contact well.	a) Rarely makes eye contact well.
8. Body language	a) In most cases employs the body language necessary for the engaging and persuasive dialogue.	a) Usually employs the body language necessary for the engaging and persuasive dialogue.	a) Sometimes employs the body language necessary for the engaging and persuasive dialogue.	a) Occasionally employs the body language necessary for the engaging and persuasive dialogue.	a) Rarely employs the body language necessary for the engaging and persuasive dialogue.
9. Phrases/ Linkers	a) In most cases employs the linking words and suitable phrases needed for communication.	a) Usually employs the linking words and suitable phrases needed for communication.	a) Sometimes employs the linking words and suitable phrases needed for communication.	a) Occasionally employs the linking words and suitable phrases needed for communication.	a) Rarely employs the linking words and suitable phrases needed for communication.
10. Task Achievement/ Completion	a) In most cases includes all of the task's prerequisites.	a) Usually includes all of the task's prerequisites.	a) Sometimes includes all of the task's prerequisites.	a) Occasionally includes all of the task's prerequisites.	a) Rarely includes all of the task's prerequisites.

5. DISCUSSION

Measuring speaking, especially in a second or foreign language, is the most challenging of the four macro skills of a language. Speaking test validity issues have significant real-world implications (Iwashita et al., 2008). According to Pennington (1999), the lack of sound grounded theory and pedagogy is the cause of the many approaches and focuses used when assessing students' speaking ability. Test results are given for speaking ability in "relevant context" since test takers want to get employment (Bradley, 1989). As a result, creating the test's construct validity was extremely difficult. To assess the contextual speaking skills, this article divided the text structure into two sections: meeting and greeting, and making the sale. The hardest part was really creating construct validity, or how the test looks, to assess test takers' confidence and pertinent contextual accuracy. The hardest part was preparing for the test. The areas of difficulty in test design were test duration, the number of interviewers, the test's target domain, CEFR level, and assessment criteria. When creating a test, test time is a crucial consideration. A 15-minute interview for a single candidate is dependable and practical for the interviewers, the candidates, and the business. Another issue with test design was how this article would start the test. This article concluded that the test will start with a "meet and greet" session after careful consideration. This is an important session since it will make it easier for the interviewers to choose the best applicant if the candidate can begin in a warm and gentle environment. Hughes (2003) refers to it as a "fresh start." Interviewers will therefore be encouraged to make a second try at expressing their thoughts rather than discouraged. Since this study involves money, the number of testers or interviewers was another difficult factor. In certain circumstances, one marker could be biased. Once more, it is extremely difficult for a single interviewer to continuously monitor the candidate's performance. As a result, this article has chosen to provide two interviewers and blind markers who will both ask the candidates questions and assign individual grades. The examinee's grade and mark will be determined by averaging the two markers' scores after the exam is over. When creating the test, this study considered the

importance of training the interviewer. In real-world experience, this study has observed interviewers who talk a lot and prevent candidates from speaking further; some begin nodding in favour of or against the candidate during the interview; some begin making notes about the candidate's performance; some remind the candidate that their performance will be rated either favorably or unfavorably; and some begin enticing candidates to give lengthy or repeated explanations of something they have misunderstood. Hughes (2003) affirms:

> "Interviewing successfully is by no means simple, and not everyone is very good at it. Interviewers must be compassionate, adaptable, and proficient in the language."

As a result, this research suggests using two qualified interviewers and markers for the SP speaking test, which was also quite difficult. Nonetheless, this article shows confidence in addressing the validity and reliability concerns at the time of test finalization.

After evaluating the existing research, this article redesigns the evaluation criteria, test construction, reliability, security, and face validity. Initially, this article created a test with a limited scope of application for hiring marketing professionals. This article has revised the test to be applicable to all types of businesses after examining the literature. Businesses fall into two categories: those that sell goods and equipment and those that offer services to customers. This test is now relevant to businesses that offer services to customers as well as those who sell goods and equipment. Second, this study created the test in three pieces with a 30-minute time constraint prior to peer review. This research has been divided into two portions, and candidates will have a 15-minute time constraint, per existing literature. All candidates will be asked a few standard introductory questions during the initial, five-minute "meeting and greeting" phase. The test-taker will next be given the information they need about the company's product or service and given three minutes to prepare. Lastly, the test-taker will present the product or service in seven different settings to the two interviewers. Along with fluency, pronunciation, and grammatical accuracy, the test-takers' body language, volume of speech, and eye contact will all be noted. Thirdly, the purpose of this study was to record audio and video to ensure the test's security and dependability. This study has decreased video recording since the review. Because of the response, this article has realized that video recording will be costly and time-consuming, and it will also put test takers in a frightening situation. According to this study, the candidate will receive the audio recording upon application by the appropriate authorities in order to increase the face validity of this article. Giving the candidate access to an audio recording could compromise the test's overall security and face validity. Therefore, the notion of giving test takers access to the audio recording was abandoned in order to improve test security. Lastly, in response to the criticism, this article changed the total marking from 100 to 50 and added a crucial criterion called "task achievement or completion" to the evaluation criteria. According to Hughes (2003), the purpose of an SP speaking test should be to evaluate "the ability to interact successfully in that language, that this involves comprehension as well as production." According to Bygate (1987), candidates should demonstrate mastery in both "informational and interactional skills" and "skills in managing interactions" when taking a speaking exam. As a result, the test in this study was created with interactional and informational skills in mind. One could debate whether or not the SP speaking test's design procedures have anything to do with validity and reliability.

The validity and reliability criteria form the foundation of this test design. Criterion validity, construct and content validity, face validity, and reliability are the criteria used to evaluate tests.

This test design continues to use the same questions for every examinee and the same product or service during the same-day interview. This has to do with face validity, which refers to giving each examinee the same treatment. In order to preserve the exam's CEFR level, this test design necessitates asking simple questions throughout the "meeting and greeting" portion. This indicates that if the interviewer asks questions on marketing core, the CEFR level will be A2; but, if the interviewer asks generic questions like "introduce yourself," the CEFR level will be B2. A classmate's input is a crucial part of construct validity when it comes to assessment criteria.

6. CONCLUSION

In conclusion, it may be concluded from the explanation above that testing can never be correct due to the repercussions of "uncertainty" in test design (Spolsky, 1995). Nonetheless, this study has made every effort to minimize test design flaws. According to Spolsky (1995), test designers can increase the test's validity by pre-testing and by examining and contrasting the test's duration with others. The exam could not be piloted in a short amount of time for this investigation. On the contrary, this study has designed the test carefully by putting comprehensible items to make the test more useful and valid. In actuality, creating a test, especially an SP speaking test, is a challenging undertaking. This article must be mindful of the various test stakeholders from start to finish, including test takers, interviewers, companies, and test designers. The study also carefully considered test backwash. In addition, linguists and researchers are testing tests to ensure that they meet all validity and reliability requirements. Every test will contain a few issues, according to this article. In this sense, the test created for this article is unique in terms of validity and reliability and has high stakes. Nonetheless, this study has made every effort to create a suitable exam that will encourage linguists and other test designers to investigate the validity and reliability of the SP speaking test for recruiting.

7. REFERENCES

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests* (Vol. 1). Oxford University Press.

Breiner-Sanders, K. E., Lowe, P., Miles, J., & Swender, E. (2000). ACTFL Proficiency Guidelines-Speaking (Revised 1999). *Foreign Language Annals*, *33*(1), 13.

Brindley, G. (1998). Describing Language Development? Rating Scales and SLA. *Interfaces Between Second Language Acquisition and Language Testing Research*, 112-140.

Bygate, M. (1987). *Speaking*. Oxford University Press.

Davidson, S. (2018). *How Valid Are Domain Experts' Judgements of Workplace Communication? Implications for Setting Standards on the Occupational English Test (OET) Writing Sub-test* (Doctoral dissertation, The University of Melbourne).

Demerouti, E., Bakker, A. B., Nachreiner, F., & Schaufeli, W. B. (2001). The Job Demands-resources Model of Burnout. *Journal of Applied Psychology*, *86*(3), 499.

Fulcher, G., & Reiter, R. M. (2003). Task Difficulty in Speaking Tests. *Language Testing*, *20*(3), 321-344.

Galaczi, E. D. (2008). Peer–peer Interaction in a Speaking Test: The Case of the First Certificate in English Examination. *Language Assessment Quarterly*, *5*(2), 89-119.

Galaczi, E. D. (2013). Interactional Competence Across Proficiency Levels: How Do Learners Manage Interaction in Paired Speaking Tests? *Applied Linguistics*, *35*(5), 553-574.

Ginting, R. S., Dalimunte, A. A., Dalimunte, M., Kurniati, E. Y., & Adelita, D. (2023). A Critical Review of IELTS Speaking Test. *JL3T (Journal of Linguistics, Literature and Language Teaching)*, *9*(2), 138-155.

Gong, K. (2023). Challenges and Opportunities for Spoken English Learning and Instruction Brought by Automated Speech Scoring in Large-scale Speaking Tests: A Mixed-method Investigation into the Washback of Speech Rater in TOEFL iBT. *Asian-Pacific Journal of Second and Foreign Language Education*, *8*(1), 25.

Hatipoğlu, Ç. (2021). Testing and Assessment of Speaking Skills, Test Task Types, and Sample Test Items. *Language Assessment and Test Preparation in English as a Foreign Language (EFL) Education*, 119-173.

Hughes, A. (2020). *Testing for Language Teachers*. Cambridge University Press.

Iwashita, N., Brown, A., McNamara, T., & O'hagan, S. (2008). Assessed Levels of Second Language Speaking Proficiency: How Distinct? *Applied Linguistics*, *29*(1), 24-49.

Jung Youn, S. (2023). Test Design and Validity Evidence of Interactive Speaking Assessment in the Era of Emerging Technologies. *Language Testing*, *40*(1), 54-60.

Khabbazbashi, N., Nakatsuhara, F., Inoue, C., Kaplan, G., & Green, A. (2022). The Design and Validation of an Online Speaking Test for Young Learners in Uruguay: Challenges and Innovations. *International Journal of TESOL Studies*, *4*(1).

Khamkhien, A. (2010). Teaching English Speaking and English Speaking Tests in the Thai Context: A Reflection from Thai Perspective. *English language teaching*, *3*(1), 184-190.

Lumley, T. (1998). Perceptions of Language-trained Raters and Occupational Experts in a Test of Occupational English Language Proficiency. *English for Specific Purposes*, *17*(4), 347-367.

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing Willingness to Communicate in a L2: A Situational Model of L2 Confidence and Affiliation. *The Modern Language Journal*, *82*(4), 545-562.

Nation, I. S. (2008). *Teaching ESL/EFL Reading and Writing*. Routledge.

Orr, M. (2002). The FCE Speaking Test: Using Rater Reports to Help Interpret Test Scores. *System*, *30*(2), 143-154.

Pennington, M. C. (1999). Computer-aided Pronunciation Pedagogy: Promise, Limitations, Directions. *Computer Assisted Language Learning*, *12*(5), 427-440.

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing*. Oxford University Press

Zhang, X., & Head, K. (2010). Dealing with Learner Reticence in the Speaking Class. *ELT Journal*, *64*(1), 1-9.