J. King Saud Univ. Vol. 14, *Lang. & Transl.*, pp 115 – 138 (A. H. 1422 / 2002)

Scaling Descriptors for Oral Proficiency Assessment in the Royal Saudi Airforce

Ghurmallah A. Al-Ghamdi

Assistant Professor,
King Faisal Air Academy
Riyadh, Saudi Arabia

(Received 8/9/ A.H. 1421; accepted 7/1/ A.H. 1422)

Abstract. The last decade of the past millennium witnessed an increase in the use of scaling descriptors for assessing the speaking skill of foreign language learners. These descriptors aim at giving a picture on what the foreign language learners are able to do in the real life situation, through assessing their ability to speak in the intended language. The aim of this research is two folds: First, a critical review of the literature is provided. Second, scaling descriptors are constructed to be used in the evaluation of cadets' spoken English in the Royal Saudi Air Force. The last part of this research addresses inter-rater reliability among six raters who used the descriptors to evaluate the performance of a number of cadets.

1. Introduction

The past decade has seen an increase in the use of scales of language proficiency. These scales try to give a picture of what a learner at a particular level can do in the real world. Thus assessment procedures should rate interactive ability in terms meaningful to all parties involved.

Since the mid-1970s, moves towards increasing European integration have added an element of urgency to the quest. Much early research was a spin-off from the Council of Europe's "Unit-Credit System" [1] approach to the foreign language curriculum. The challenge has produced some interesting research but, to date, there is no unanimity about what constitutes an accurate, balanced and valid scale of oral proficiency. Indeed, as Alderson (1991) puts it, when outlining problems involved during the process of arriving at descriptors. He wrote: "Several problems were encountered during this process, including the obvious one of which features/criteria to include/exclude, and how to define them in the scales" [2, p. 81].

My intention is to review recent work in the field, examine the issues which stand in the way of consensus and attempt to draw up a preliminary "bank of descriptors" which could form the basis of a new Royal Saudi Air Force Test of Oral Proficiency.

2. Overview

Shohamy [3] classified oral tests into two categories: these are pre-communicative and communicative authentic. The precommunicative oral tests, mostly non-reciprocal, were only conducted in language laboratories and involved repetition of utterances or even words and/or alternating patterns of substitution drills during the course of choral or individual drilling. The communicative oral tests, on the other hand, are interactive-based which involved an interviewer, an interviewee and sometimes an independent assessor. The classification, therefore, reflected the effect of language teaching methodologies existing at the time of the inauguration of each type.

It is obvious that the audiolingual approach to language teaching is associated with the precommunicative approach to oral language testing, and the emphasis of teaching language for communicative purposes is associated with the latter. Indeed, as Davies [4, p. 136] put it "Language Testing . . . grew out of language pedagogy needs, where it still firmly belongs."

The first version of the American Foreign Service Institute (FSI) test, later known as the oral interview (OI) was an example of real communicative oral tests [5]. The test itself involved a test giver (examiner) and a test taker (examinee). The outcome of the face-to-face interview yielded a language sample, which is assessed on the basis of a rating scale, which consisted of descriptors representing different levels, from 0 to 5 with intermediate pluses and minuses.

The FSI (OI) was later adopted by the American Council on the Teaching of Foreign Languages [6] which was then developed into different proficiency levels, which range into four levels of proficiency. These are "novice", "intermediate", "advanced" and "superior". Each level included several numbers of sub-levels, such as "lower intermediate" and "upper intermediate".

In Europe, on the other hand, with the move towards integration of the European countries, the continent has witnessed a "proliferation" [7, p.217] of European scales constructed independently of the American scales. The following are examples:

a) The Euro centres scales of language proficiency [8].
b) The British National Language Standards [9].
c) The Association of Language Testers in Europe (ALTE) Framework [10].

These scales can be classified in terms of their assessment procedure into two categories: "real life" or "behavioural approach" as Bachman [11] called them because they attempt to assess the candidate's speech samples in terms of what he can do in the real world. The (ACTFL) guidelines are examples. The shift of emphasis towards "real-life" tests is a move towards authenticity in language testing.

The other kind is "interactive-ability" scales, which assess candidates' language performance in terms of the features of their speech, whether linguistic or pragmatic. The main problem with these scales or descriptors is that they can be confusing to the assessor because they are often detailed and long.

3. Uses of descriptors of language proficiency

Descriptors in language proficiency testing are used for a variety of reasons. The first is to describe levels of language performance to enable test users to make judgements about candidates. This particular use is described by Alderson [2, p. 72] as 'user oriented,' A level of performance together with its associated descriptor helps users to decide whether the candidate reached a particular level to be admitted to a job or a field of study, something which a numerical score does not always give, particularly if the validity and reliability of the test in question are unknown.

The other use of descriptors is to facilitate assessing an individual's performance, especially if the test takes the form of an interview. By comparing the behavioural outcome of the candidate against the scale of the descriptors, an assessor should be able to make a more precise judgement assuming he adheres to the descriptors while rating. This purpose has been given the term "assessor-oriented" Alderson [2, p. 73].

A third purpose is to provide the test constructor with guidelines on how to construct tests because normally descriptors give clear guidelines or specifications for the intended test in terms of what texts or tasks are more suitable for a particular level of performance.

3.1 North and Schneider's Scale: A Closer Look

In an attempt to construct coherent taxonomies of pragmatic and linguistic features that would eventually be used for assessing language proficiency, North and Schneider [7] looked at a number of existing descriptors. They then concluded that we should not wait for empirical validated research in order to build upon its conclusions. In other words, practitioners (syllabus writers as well as teachers) cannot wait for theories of language proficiency to emerge. Rather, they are required to work with taxonomies that are acceptable but with certain conditions. These conditions, according to North and Schneider [7], are:

1. Proper documentation of the experience.
2. Consensus between existing models of language proficiency.
3. Organisation of language activities.
4. Making sure that descriptors involve proper criteria of assessment.

It is necessary to be aware of the limitations of the study – as the authors say themselves, what's actually being scaled is *"the teachers' collective interpretation of* the difficulty of these descriptors, a mapping of the behaviours of one group by another group" [12,pp. 67, 223]. Not the same as objectivity but "attractive...in the context of developing a common framework."

Attractive, possibly, but not without its dangers: the history of language teaching, especially, is littered with quick fixes that have become orthodoxies.

(N.B. that "descriptors asking teachers to guess about activities (generally work-related) beyond their direct experience: telephoning; attending formal meetings; giving formal presentations; writing reports and essays; formal correspondence" showed a high degree of statistical misfit and to have their difficulty underestimated when analysed with the main construct.'. Also note the loss of items relating to 'receptive strategies' .

You also come up against a *practical* limitation here: there's no reason, presumably, why descriptors shouldn't be drawn from linguistics, psycholinguistics or psychology. The descriptor pool itself derived from a survey of 41 language proficiency scales [13], which were dismantled and definition for each scale assigned to a provisional level. Each descriptor was split up into sentences (sometimes compound) and allocated to a provisional category. His theoretical framework seems somewhat questionable, too: his overall scheme draws on consensus between existing models of communicative competence and language. (e.g., [14, p 81; 15; 11; 16;17]. Also organized language activities under the headings "reception", "interaction" and "production" [18]. See [19] for a brief account. Briefly, communicative language competence (linguistic, pragmatic, sociolinguistic) is seen as part of overall human competence. Learners draw on these competencies, depending on circumstances, and adopt appropriate strategies in order to complete specific task in language activity concerned. Product is text. (In understanding, too?) With speaking/interaction, it worked thus:

A. Communicative Language Abilities			**B. Strategies**			**C. Qualitative aspects of language proficiency**		
Recep-tion	Inter-action	Production	Recep-tion	Interaction	Production	Prag-matic	Linguistic	Socio-linguistic

Descriptors for **A** correspond to Bachman's "real-life" and Alderson's "constructor-oriented" classification. So you have things like "*Can enter unprepared into conversations on familiar topics*".

Those for **C** = Bachman's "interactive-ability" and Alderson's "assessor-oriented" classification (e.g., descriptors for *fluency*). (This is where things seem to me to get a bit messy, really: why is fluency "pragmatic" rather than "linguistic"?)

Strategies somewhere between these: so you have descriptors for things like *compensating* (e.g., defining something when you can't recall the word).

Couple of points here: to what extent are "strategies" peculiar to the L2 learner? Conceivably, a native speaker may sometimes want to 'define something when he can't recall (or doesn't know) the word.

Another thing that seems a little odd is that "qualitative aspects of language proficiency" should be defined in terms of the "text" (i.e., what the candidate *produces*) whereas "communicative language abilities" are defined in terms of what the candidate *does*.

The claim that this framework is in some sense "*a priori* construct validation in relation to theory as well as to the way users think" [20, p 21] seems to me very doubtful. Even if one accepted, for the sake of argument, that it were, I don't see that plugging descriptors derived from scales that were *not* developed within such a framework into the framework is valid. (so you have the familiar anomalies (e.g., that there is a descriptor dealing with "complaints", but not one that deals with "insults" or "snubs" (The latter is not appropriate to our situation, but is mentioned for the purposes of illustration.).
replication of the '94 study in '95 showed "very satisfactory consistency" between teachers of English (94), and teachers of English, French and German .

In their conclusion, North and Schneider cite various authors who believe that "the development of common framework scales should not be attempted before research has provided adequate empirically validated descriptions of language proficiency and of the language learning process." But, as he notes (and quotes), we can't hang around waiting for "*full-fledged theories of language proficiency to emerge from research laboratories.*"

The basic problem is the 'tension between theoretical models developed by applied linguists (which are incomplete) on the one hand and operational models developed by practitioners (which may be impoverished) on the other hand.'

How could we do this? Extensive replication may be one way: North and Schneider's work has demonstrated that consensus does exist among practitioners. What we might do is (a) accept that such consensus exists on some items but not on others (e.g., the 'new categories' (such as the interaction strategies). One fairly obvious hypothesis (obvious, indeed to the point of being banal) is that the 'new categories' are simply unfamiliar to most practitioners.

4. Arriving at descriptors for speaking skill to be used for the cadets of King Faisal Air Academy (KFAA)

4.1 Setting

Cadets at the KFAA are required to obtain a level of language proficiency that enables them to study aero sciences in the medium of English. Upon joining the KFAA, the majority of candidates, about 80%, require to be trained in the English language at various levels. Some spend 10 weeks, others spend up to 52 weeks of intensive training, but the majority of them spend between 30-35 weeks. They are trained to be pilots.

Pilot trainees, upon entry, are given a placement test. The candidate whose English is classified as 'novice' has to begin studying Book One. The course contains six main books. The series is entitled *English for Air Cadets*. The course specifications originate in the PhD research carried out by the author at Edinburgh University [21].

Once candidates complete Book Six successfully, they have to sit for an oral test. Passing this test is a prerequisite for sitting another paper-pencil test, mainly testing listening comprehension, grammatical structures and vocabulary.

However, the oral test we are now using is in need of upgrading. The main reason for this is that it is too restricted in terms of its rating system. It only looks at fluency, accuracy and range. It does not capture more important linguistic and pragmatic features. In addition, although the descriptors, are defined, their definitions are subject to different interpretations. Therefore, inter-rater reliability of scoring is questioned.

The aim of this research is to identify what our cadets are required to be able to do at the end of each book. Following this we would be to construct descriptors for assessing oral proficiency.

Below is a breakdown of the enabling skills for Books One through Four. Illustrative examples are provided for Books Three and Four.

Book One

Can use and respond to greeting expressions for different times of the day. "Hello", "How are you", etc.
Can respond to classroom instructions, e.g., "Close the door." "Stand up." "Write the answer." "Do the status board."
Can identify letters, numbers and number words.
Can identify acronyms and abbreviations.
Can ask questions about spelling.
Can ask and answer questions about himself and his family, giving personal information.
Can discuss photos of very basic information.
Can ask and answer questions about countries and nationalities.
Can make and answer yes/no questions.
Can give personal information about others, males/females.
Can ask and answer questions about job information, especially jobs in the forces.
Can identify officer ranks, NCO ranks and cadet ranks.
Can identify family relations.
Can describe people in terms of appearance, colours and clothes.
Can complete a form.
Can identify the different parts of the body.
Can describe feelings.
Can identify and talk about objects of basic nature.

Book Two
Can describe places: inside buildings, offices and hotels using appropriate prepositions.
Can make sentences about the location of objects.
Can complete and talk about a plan or map.
Can describe shapes: square, circle, etc.
Can ask and answer questions about directions.
Can describe towns.
Can find the differences between two pictures.
Can ask and answer questions about ability.
Can take basic notes to complete a table about different people's ability.
Can tell the time.
Can express likes and dislikes: food, drinks, and favourite games.
Can discuss activities: sports, travel, etc.

4.2 EAC BOOK THREE. CADET PROFILE.

PERSONAL INFORMATION:

Present Activities

1a. Daily routines

What do you do on the weekends?
I ... ...
Describe your routine in the academy.
First, I ... . Then ... . Finally, ...

1b. Pastimes

Do you play any sports?
Yes, I do. I play football.

Who do you play with?
I play with my friends at the stadium. What about you?/Do you play football?
No, I don't play football, but I like watching it on the TV.

FACTUAL INFORMATION:

Exchange information by referring to a map and can use compass points

Which direction is?
It's north/south/east/west.
... north west/east ... south west/east.

To and from the north ...

Where is?
It's in the north.
It's on the west coast

Can exchange information about measurements

i. Fractions
¼ = quarter, ½ = half, ¾ = three quarters

ii. Decimals
5.05 = five point zero five

Does Sami play any sports?
Yes, he does. He plays football.
Where does he play?
He plays at the stadium.

Past activities

What did you do last Friday?
I had a picnic.
Where did you go?
I went to Thumamah with some friends. What about you?/What did you do?
I stayed at home with my family.

Future plans

What are you going to do next week?
I'm going to Dhahran with some friends.
Are you going to the beach?
Yes, we are. What about you?/What are you going to do?
I'm staying at home with my family.

iii. Speed
How fast does it go?
130 kph

iv. Weight
How heavy are you?
I'm about 62 kilos.
How heavy is it?
It's exactly 76.5 kilos.
What does it weigh?
62 grams.

v. Countable and uncountable nouns
How much rice do you have?
About 2 kilos.
How many oranges do you have?
Five.

EAC BOOK THREE. CADET PROFILE.

You need to understand and use the following expressions in the classroom.

CLASSROOM INTERACTIONS:

1. Can understand and obey classroom instructions.

Open your books.
Don't speak Arabic.
Stop writing now.

2. Can ask for things with a reason.

Can you give/pass/lend me a pen, please? I don't have one.
Yes, of course, here you are.
No, I'm sorry. I don't have one.
Can I borrow your ruler, please? Mine's at home.
Do you have an eraser/rubber?

3. Can ask for permission with a reason.

I'm sick. Can I go to the clinic, please?
Yes, of course./No, you can't.
Can I close the door, please? It's hot in here.

4. Can apologise with a reason.

Why are you late?
I'm sorry I'm late. I had to see the colonel.
Where's your pen?
I forgot to bring a pen. I'm sorry.

5. Can express how he feels in simple terms.

How do you feel?
Oh, not too bad./Okay.
I feel very sick today. I have a bad headache.
What's wrong with you?
I'm not very happy.
Why?
I can't go home this weekend

PERSONAL INFORMATION

1a. Can ask and answer questions and provide personal information about selves and other people.

Where were you born?/was Sami born?
How tall are you?/is he?
How much do you weigh?/does he weigh?

I'm 22 and I was born in Riyadh. I'm 1.60m tall and I weigh 90 kg. My friend, Sami, is 20 years old and he was born in Dhahran. He's 1.55 m tall and he weighs 82 kg. So, I'm taller and heavier than him. Ali is 1.75 m tall and weighs 95 kg. He's the tallest and the heaviest.

1b. Can answer 'why' questions: explaining a problem or giving a reason.

Why are you running?
Because I'm late for class.
Why's Fahad looking out the window?
He's thinking of his family.

Why can't he ride that horse?
He's too heavy.
Why didn't you buy that Corvette?

Tell me something about your hometown.
I'm from Abha. Abha is in Asir in the southwest of Saudi Arabia. It's a beautiful city in the
mountains. Abh a is about 1,000 km from Riyadh, and it takes about an hour and a half half to get to Riyadh by air.

What is the weather like in Jeddah?
In *Jeddah in the spring, it is cool and it Sometimes rains.*
It gets hotter in the summer and temperatures reach about 40 degrees centigrade. It is also very humid.

Can ask and answer questions and describe people and activities in a picture.
Describe the picture.
What is the man on the left doing?
There are two men in this picture. They are both wearing robes. One man is driving a Toyota and the other is riding a camel.
All the men in this photograph are playing football. Some of them are wearing blue; others are wearing red. One of them is standing; the others are running. I can see a a man holding a red flag. Everyone is looking at him.

FACTUAL INFORMATION

Can ask and answer questions about planes and cars.
How long is the Jet stream 91?
It's 19.25 metres long.
How wide is it?
18.24 metres.
How high is it?
5.74 metres high.
What's its weight?
It weighs 6,416 kg.
What's its top speed?/How fast can it fly?

I'm not rich enough.
It was too expensive.

1c. Can talk about personal experiences.

What did you do last weekend?
I went to Jeddah.

Why did you go to Jeddah?
I visited some friends...

2. Can ask and answer questions and describe their hometown and climate.

Where are you from?
Where's that?
What's it like?
How far is it from Riyadh to xxx?
How long does it take to get from Riyadh to xxx?

546 k/h. It can fly at a maximum speed of 46 k/h.

ask for and give instructions using sequencers.

How do you change a wheel on a car?
First loosen the wheel nuts. Next ... . r that . Then ... . Finally ... .

EAC BOOK FOUR. CADET PROFILE.

You need to understand and use the following expressions in the classroom.

CLASSROOM INTERACTIONS:

Can express feelings.

What's wrong with you.
I'm not very happy.
Why?
I can't go home this weekend.

Can ask and answer questions about activities and respond to the situation in the classroom.

Stop writing on the desk.
I'm not writing on the desk. I'm writing in my book.

Can answer questions about the weather.
What's it like outside?
It's raining..
Can understand, obey and give orders.

Cadets mustn't smoke in the academy.
You must salute an officer.

Can use 'like' to ask and answer questions about people.

What does he look like?
He's tall and thin.

What's he like?
He's clever and funny.

5. Descriptors for an oral test

Having drafted a set of descriptors for the first four books in terms of what is expected of cadets to be able to do at the end of each book, we are now going to relate that to North and Schneider's 1998 studies. We will examine their scale in terms of relevance and cultural appropriateness to our situation.

It is my understanding that once these descriptors are compiled, then it is necessary to show them to teachers and experienced markers for comments and modifications. However, the following descriptors were compiled on the basis of my own experience in the field of language testing, the language curricula and of the King Faisal Air Academy's context. Also, the list drawn by North and Schneider (1998) was made use of after redundant and irrelevant descriptors were eliminated.

6. Preliminary analysis for the exit test

6.1 Purpose of test

1. To make inferences about pilot trainees' speaking ability in general English contexts.
2. To make decisions about the amount of time they need in order to reach a satisfactory level of performance.

3. To be used for diagnostic purposes to identify areas of weakness.

6.2 Domain

Language instructional.

6.3 Tasks

The tasks will be identified by analysing the content of the language course. The domain for these students is broadly identified on the basis of the types of tasks they are going to deal with, including carrying out conversation with their instructors, responding to and asking questions, giving briefings, describing in a narrative form. Hence role-playing, re-telling of events and telling the differences.

6.4 Test takers

Test takers are males and 18 years old, mainly Saudis. Small numbers come from the GCC countries and some from African countries: Egypt, Senegal, Djibouti, Algeria and Sudan. They come from a variety of social classes but all except African cadets have the same educational background.

6.5 Language ability

Candidates will have successfully completed six books of the course entitled *English for Air Cadets* According to their proficiency level upon entry, they will have spent between 10 and 35 weeks at the time of taking the test.

6.6 Test construct

Linguistic and pragmatic definitions of language ability (see the assessment scheme below).

Grammatical Accuracy

Level	Description
Advanced	The candidate uses lexis beyond the core level correctly and employs clausal and inter-clausal patterns with consistent inflectional and syntactic accuracy.
Intermediate	The candidate uses core lexis correctly and employs familiar clausal and supra clausal patterns with inflectional and syntactic accuracy with variable consistency, but shows awareness of the need for such accuracy by self-correction and repair.
Novice	The candidate uses basic learned clausal patterns with frequent inflectional and syntactic errors and omissions.

Fluency

Level	Description
Advanced	The candidate interacts with a degree of fluency and flexibility that allows for interaction with a native speaker without imposing strain on either party.
Intermediate	The candidate can produce discourse above the level of the clause, on familiar topics and with evident pauses for grammatical and lexical planning and "repairs" and corrections, especially in longer stretches of discourse.
Novice	The candidate operates only at the level of the clause, within learned topics and using learned patterns, and with frequent breakdowns

Information Exchange

Level	Description
Advanced	The candidate can ask and answer relevant questions concerning factual information derived from unfamiliar forms, and make relevant inferences using this information
Intermediate	The candidate can ask and answer relevant questions concerning the main points of factual information derived from familiar forms (e.g., graphs, pictures, tables, diagram).
Novice	The candidate can ask and answer relevant questions concerning basic personal information (e.g., name, date of birth, etc.)derived from learned forms(e.g., immigration forms, hotel registration forms , etc.).

Describing and Narrating

Level	Description
Advanced	The candidate can give basic details of unfamiliar event(e.g., an accident) if allowed the time to prepare himself.
Intermediate	The candidate can describe familiar habits,routines,likes and dislikes.
Novice	The candidate can describe people(e.g., their relationship to him, their age, occupations etc.)in simple learned terms.

These analytical grids are combined to yield the three-grade assessment profile detailed overleaf. These are *holistic* scales which describe a candidates overall behavior.

Advanced

The candidate can communicate information on unfamiliar topics without assistance from the interlocutor. He can understand and respond to the main points of short stretches of discourse, both on familiar and unfamiliar topics, and is capable of making relevant inferences beyond the text itself. He can give the basic details of an unfamiliar event (e.g., an accident) if allowed the time to prepare himself. He can ask and answer relevant questions concerning factual information derived from unfamiliar forms, and make relevant inferences using this information. He uses lexis beyond the core level correctly and employs clausal and inter-clausal patterns with consistent inflectional and syntactic accuracy. He interacts with a degree of fluency and flexibility that allows for interaction with a native speaker without imposing strain on either party.

Intermediate

The candidate can communicate information on familiar topics, with limited assistance from the interlocutor. He can understand and respond to the main points of short stretches of discourse on familiar topics, but is only capable of making very limited inferences beyond the text itself. He can describe familiar habits, routines, likes and dislikes. He can ask and answer relevant questions concerning the main points of factual information derived from familiar forms (e.g., graphs, pictures, tables, diagrams). He uses core lexis correctly and employs familiar clausal and supra-clausal patterns with inflectional and syntactic accuracy with variable consistency, but

shows awareness of the need for such accuracy by self-correction and repair. He can produce discourse above the level of the clause, on familiar topics and with evident pauses for grammatical and lexical planning and 'repairs' and self-corrections, especially in longer stretches of discourse.

Novice

The candidate can communicate basic information on learned topics, with extensive assistance from the interlocutor. He can understand and respond to basic (i.e., clausal) questions and instructions addressed to him, with frequent repetitions. He can describe people (e.g., their relationship to him, their age, occupations etc.) in simple learned terms. He can ask and answer relevant questions concerning basic personal information (e.g., name, date of birth etc.) derived from learned forms (e.g., immigration forms, hotel registration forms etc.). He uses basic learned clausal patterns with frequent inflectional and syntactic errors and omissions. He operates only at the level of the clause, within learned topics and using learned patterns, and with frequent breakdowns.

7. Arguments

It can be noticed that the descriptors were confined to three levels of performance. These are: "advanced user", "intermediate" and "novice". The reason is that, upon entry, cadets are classified into three levels of performance according to the following percentages:

1- 17% can be classified as "upper intermediate".
2- 30% can be classified as "lower intermediate".
3- The remainder, who represent the majority of about 53%, are false beginners.

These percentages are categorized on the basis of a placement test given to cadets upon entry.

The course, which is being taught at the moment, prepares cadets to the "upper intermediate" level of performance. It is also assumed that those whose background is "upper intermediate" level upon entry, advance their level of performance; through the advanced modules that they are going to study.

The other main reason is that too long scales, regardless of their appropriate content, confuse the assessors, particularly if the candidate is to be assessed on the spot.

The test I envisage should represent more than one single speech style, i.e., of the interview. It has to have different tasks to be considered as a valid measure of language proficiency.

In addition, it should serve as a positive wash back effect to language teaching; that is, it should emphasise the practice of speaking in the classroom. It should also be reliable, i.e., consistent in terms of reliability of marking (inter-rater reliability) and consistent in terms of tasks given to candidates.

7.1 Test Tasks

In order to elicit oral performance, cadets will be given a set of tasks. Each will be assessed independently in terms of linguistic and pragmatic features, as illustrated in the scale discussed above.

Task 1: Telling the difference between two pictures

A cadet is presented with two pictures, which differ in some aspects. He is required, in a narrative form, to tell the differences between the pictures, e.g., the pictures could be about two airfields which differ in terms of the location of the tower, number of runways, type of and number of aircraft, ramps, etc., or they could be of a general nature: two rooms with different objects, for example.

Task 2: Role-play

This section includes a variety of speech functions in different sociolinguistic/situational contexts, e.g., booking, car hiring, etc. The candidate is required to carry out a conversation with the examiner on one of these or similar topics.

Task 3: Story telling

Candidates are presented with a three-minute segment video. The video is then switched off and the candidate is asked to give an account of the events of the segment.

Topic discussion, on the other hand, was avoided on the grounds that potential topics such as education, accommodation, travel, media, etc., might be memorised and re-delivered. Topics like these are finite and it may well be the case that cadets, being foreign language learners, memorize them. Therefore, the assessor might be confused in the sense that he might attribute the fluency of the delivery to the actual level of performance, whilst it might be attributed to memorization.

8. Summary of the overall specifications for the test

Objective:	To determine whether trainee pilots are able to speak the English language in the context of general English. It is assumed that specialised vocabulary items will be acquired at a later stage.
Answering procedures:	The candidate will speak to a native speaker or to a non-native speaker of native speaker ability on a number of tasks in the presence of an independent assessor.
Number of tasks:	Three: Comparing and telling the differences

	between two pictures, re-telling the story of a film segment and a role-play.
Time:	Six minutes for each task and two minutes for relaxation purposes.
Rating scale:	A three-point scale 1-3, a number of linguistic and pragmatic tasks such as fluency, accuracy, overal interaction, information exchange, turn taking describing and narrating.

8.1 Bands and scores

As can be seen from the scales, they should help assessors to rate candidates elicited performance, provided assessors adhered to them. One point here, it might look as if the bands are too restricted but the problem is that we are dealing with a skewed population. Most of our cadets are band 2 and below. But, one advantage of the adopted scale is that it captures many linguistic and pragmatic features, the advantage of this being it could produce more reliable ratings. In this respect, Alderson (1991) wrote: "In order to produce reliable ratings, it is often necessary to provide a series of rating scales, each focussing on a different aspect or combination of aspects of the performance."

8.2 Rating

Because the features that capture performance are many and detailed, it would be extremely difficult for the interlocutor to assess elicited spoken performance while conducting the test. The solution to the problem is either to record the test and later listen to it and rate it, or to have an independent assessor who would assess performance on the spot, but has nothing to do with conducting the test. Both solutions are expensive but the latter is more convenient.

The assessor may take notes during the course of the test of the candidate's performance and later relate it to each point on the scale. It could be the case that candidates may perform differently at different features of the scale. Therefore, the final band score would be the aggregate of the different points as follows:

Advanced user	3
Intermediate user	2
Novice user	1

The problem with this is that band scores might yield minuses and pluses points. This can be overcome by approximating the bands or moderation.

9. Piloting the test and training assessors

Regardless of how well the test was constructed, the best way to decide whether the test is actually working is by field-testing it in order to decide on its practicality, appropriateness of content and difficulty of tasks or items. In other words, we must establish its validity and reliability in advance.

In this respect, Douglas [22] quotes Alderson [23]: "Alderson reports on three different studies on which both experienced teachers and professional testers could not agree on what skills were being tested or how performance would be assessed."

In our case, the standard fieldwork procedures will be applied. First, the test content and the assessment procedures will be shown to colleagues for initial comment on its content, associated rating scales, appropriateness of rubrics and clarity of instructions. Assuming consensus, the test will eventually be given in its final draft to a representative sample of the population to check on its validity and reliability.

On the other issue, i.e., training of assessors, it is of paramount importance that performance tests, such as this test, should be conducted only by those who are adequately trained in order to ensure inter- and intra-rater reliability.

First, the scale has to be explained thoroughly to them. They should also be enlightened of tasks' content as well as proper elicitation techniques. Then, a check on their ability to sustain consistency of rating is necessary. The latter is done by observing them actually conducting the test itself. Second, data will be collected from the trials and computation of reliability should be carried out. It is also well known in the literature of language testing that adhering to the scoring procedures will make the test semi-objective, since no matter what, performance tests will remain largely subjective.

Continuous assessment of cadets' oral proficiency

So far, we have confined our discussion to the "exit test": the oral proficiency test given at the end of the course. We said earlier that from an educational point of view the test should opt for promoting the practice of speaking skills in the classroom. Assuming that the tests themselves exert some kind of influence on the teaching practices in the classroom, teachers should give oral skills priority in everyday teaching practices without, of course, neglecting other aspects of language teaching. The technique of assessing cadets' oral performance should initially be confined to tasks reflected in each book. We are now going to draw up some themes that should be elicited by the teacher at the end of each book. They are not, by any means, exhaustive.

Book One

Book One prepares cadets to use basic social language, give basic personal information about oneself as well as others, talk about objects, give factual information and obey classroom instructions. It also prepares then to ask for repetition, clarification of meaning.

Book Two

Book Two continues to prepare cadets on how to give extended personal information, how to talk about places, give directions and provide factual information.

Book Three

Book Three prepares cadets on how to ask and answer questions for the purpose of how to develop conversation, how to talk about present time and past time activities, future plans, ability to exchange information using a map and compass points. It also prepares him on how to exchange information about measurements as well as classroom interaction.

Book Four

Book Four prepares cadets on how to give extended descriptions of people, things and objects. It also prepares them on how to reason using basic ESP related materials. It is well established in the literature [24; 25] that instructors would develop a broad and sound understanding of their students' achievements if they kept testing them. Repeated testing for achievement purposes is also a means of assessing language teaching pedagogy. However, assessment of the above-mentioned tasks should take one of the following (a) the candidate is able to perform the required tasks or (6) he is not able (not ready yet). At this stage these books should prepare cadets on how to carry out a structured oral interview.

For **Book One**, he can be asked to do the following tasks:

Stage 1:	How to respond to greetings and orders.
Stage 2:	What's your name, military number? How old are you? What book are you doing? etc.
Stage 3:	Talking about basic objects: car, aeroplane (using pictures as stimuli).
Book Two	
Stage 1:	To be given pictures of people for the purpose of describing their appearance.
Stage 2:	Using pictures as stimuli to talk about places.
Stage 3:	Using plans to give directions.
Book Three	
Stage 1:	Talk about his daily routines.
Stage 2:	Talk about his past activities and future plans.
Stage 3:	Ability to exchange information by referring to a map.

Book Four

Stage 1:	Exchange information about measurements to talk about speed and weight, using decimals, fractions.
Stage 2:	Describe this picture.
Stage 3:	Presentation of a topic.

10. Results

10.1 Introduction

One of the main problems of performance tests, both written and spoken, is reliability of scoring. In order to test inter-rater reliability with the new interview, a number of interviews were recorded on videotape. These were later played and six assessors, all except myself are native speakers of English, were asked to rate subjects' performance on three scales, overall interaction, lexical and grammatical accuracy and fluency.

Scale 1: overall interaction

Table 1 shows correlations between ratings for the first scale (overall interaction):

Table 1. Correlations between raters on overall interaction using Spearman's Rho.

	G	KR	CT	GR	YN	PE
G		1.00**	0.892*	1.00**	0.750	0.949 **
KR			0.892*	1.00**	0.750	0.949 **
CT				0.892*	0.649	0.872
GR					0.750	0.949 *
YN						0.632
PE						

In terms of inter-rater reliability, this results in a score of 1.09, which in turn yields a Pearson correlation between 0.79 and 0.80 (which is significant). However, with further refinement of the descriptors and further training of the raters, a much higher correlation should be possible [26, p 535]

Scale 2: grammatical accuracy

Table 2 shows correlations between ratings for the second scale (grammatical Accuracy):

Table 2. Correlations between raters on grammatical accuracy using Spearman's Rho.

	G	KR	CT	GR	YN	PE
G		0.186	.395	.363	-0.395	0.363
KR			0.825	0.649	0.825	0.973**
CT				0.918*	0.500	0.918*
GR					0.344	0.763
YN						0.688
PE						

In terms of inter-rater reliability, this yields a score of 0.95, which in turn yields a Pearson correlation of 0.74. Once again, further training is clearly necessary.

Scale 3: fluency

Table 3 shows correlations between ratings for the third scale (fluency):

Table 3. Correlations between raters on fluency using Spearman's Rho.

	G	KR	CT	GR	YN	PE
G		0.917*	0.648	1.00**	0.649	0.973**
KR			0.884*	0.917*	0.811	0.892*
CT				0.648	0.688	0.631
GR					0.649	0.973**
YN						0.632
PE						

In terms of inter-rater reliability, this yields a score of 1.09, which in turn yields a Pearson correlation between 0.79 and 0.80 (which is significant).

11. Conclusion

Scoring performance tests and maintaining perfect inter-rater reliability continue to be a problem and difficult to achieve. However within the limitations of this study, it has been proven that there is a great deal of unanimity among the six raters who assessed candidates oral performance according to the scaling descriptors set out above. Significant correlations coefficients were obtained using Spearmen's RHO after correction for attenuation. However it is possible to improve inter-rater reliability by more training as has been suggested above. Attempting subjective tests should not be avoided but further research is required particularly on ways of improving writing of descriptors.

References

[1] Van Ek, J. *The Threshold Level in a Unit/Credit System.* ed. Trim, et al., (1980), 89-146.

[2] Alderson, J. Charles. "Bands and Scores" In *Language Testing in the 1990's.* (1991), 71-86.

[3] Shohamy, Elana. "The Validity of Direct Versus Semi-direct Oral Tests." In *Language Testing.* Vol. 11 (1), (1994), 99-122.

[4] Davies, Alan. "Language Testing in the 1990's." In *Language Testing in the 1990's.* (1991), 136-149.

[5] Jones, R. A. "Testing: A Vital Connection." In *The Language Connection from the Classroom to the World.* Ed. J. Phillips. The ACTFL Review of Foreign Language Education. Vol. 9. Skokie, IL: National Textbook Co., (1977, 237-265.

[6] American Council on the Teaching of Foreign Languages. *ACTFL Proficiency Guidelines.* New York: American Council on the Teaching of Foreign Languages (1986).
[7] North, B and Schneider, G "Scaling Descriptors for Language Proficiency Scales." In *Language Testing.* Vol. 15 (2), (1998), 217-263.
[8] North, B. *Standardisation of Continuous Assessment Grades* (1991). Strasbourg: Council of Europe.
[9] Languages Lead Body. *National Standards for Languages: Units of Competence and Assessment Guidance.* UK Languages Lead Body, July, 1992.
[10] *Association of Language Testers in Europe. ALTE Handbook of European Language Examinations and Examination System: Descriptions of Examination Offered and Examinations Administered by Members of the Association of Language Testers in Europe.* Cambridge: University of Cambridge Local Examinations Syndicate, (1998).
[11] Bachman, L. F. *Fundamental Considerations in Language Testing.* Oxford: Oxford University Press (1990).
[12] Pienemann, M. and Johnston, M. "Factors influencing the development of Language Proficiency." In *Applying Second Language Acquisition Research.* Adelaide: National Curriculum Resource Centre, (1987), 89-94.
[13] North, B. *Scales of Language Proficiency: A Survey of Some Systems.* Strasbourg: Council of Europe (1994).
[14] Canale, M and Swain, M. "Theoretical Bases of Communicative Approaches to Second Language Teaching and Testing." *Applied Linguistics,*1 (1980), 1-47.
[15] Van Ek, J "Objectives for Foreign Language Teaching." In *Scope.* Vol 1. Strasbourg: Council of Europe (1986).
[16] Skehan, P. "Analysability, Accessibility and Ability for Use." In *Principles and Practice in Applied Linguistics.* 5. Ed. G. Cook and S, Seidehofer. Oxford: Oxford University Press (1995).
[17] McNamara, T. "Modelling Performance: Opening Pandora's Box." *Applied Linguistics,*16 (1995), 159-179.
[18] Brumfit, C.J. "Concepts and Categories in Language Teaching Methodology."*AILA Review,*4(1987), 25-31.
[19] North, B. "Perspectives on Language Proficiency and Aspects of Competence." In *Language Teaching,* 30, (1997) 93-100.
[20] Griffin, P. E. and McKay, "Assessment and Reporting in the ESL Language and Literacy in Schools Project." In *National Languages and Literacy Institute of Australia,* ESL Development. Vol. II (1992).
[21] Al-Ghamdi, G. *English Proficiency in the Saudi Air Academy: Validating a New Test Battery.* Unpublished Ph.D. Thesis University of Edinburgh, (1987).
[22] Douglas, Dan "Assessing Languages for Specific Purposes." *Cambridge Language Assessment Series.* Ed. J. Charles Alderson and Lyle Bachman. Cambridge: Cambridge University Press (2000).
[23] Alderson, J. Charles. "Judgements in Language Testing." In *A New Decade of Language Testing Research.* Ed. D. Douglas and C. Chapelle. Arlington, VA: TESOL, (1993), 248-259.
[24] Gipps, C. V. *Beyond Testing: Towards a Theory of Educational Assessment.* London: Falmer Press, (1994).
[25] Teasdale, Alex. and Leung, Constant. "Teacher Assessment and Psychometric Theory: A Case of Paradigm Crossing?" In *Language Testing.* 17, (2000), 163-184.
[26] Hatch, E and Lazaraton, A. *Design and Statistics for Applied Linguistics,* Newbury House, (1991).

المعايير الوصفية المقننة لتقويم الكفاية اللغوية

غرم الله عبد الله الغامدي

أستاذ مساعد، كلية الملك فيصل الجوية

الرياض، المملكة العربية السعودية

ملخص البحث. شهد العقد الأخير من القرن الماضي زيادة في استخدام المعايير الوصفية المقننة لتقويم الكفاية اللغوية لمهـــارة الـــتحدث عـــند متعلمي اللغة الإنجليزية بوصفها لغة أجنبية وكذا اللغات الأجنبية الأخرى. وهذه المعايير الوصفية تهدف إلى إعطاء صورة عن ماذا يمكن لهذا المتعلم أن يؤدي في الواقع من خلال تقويم قدرته على التحدث بـــتلك اللغة. إن الهدف من هذا البحث هو مراجعة نقدية لأدبيات هذا المجال ووضع معايير وصفية لتكون أساسا تســـتخدم في تقـــويم مهارة التحدث عند متعلمي اللغة الإنجليزية في القوات الجوية. كما أن الجزء الأخير من هذا البحـــث يهدف إلى دراسة درجة الثبات عند ستة من الممتحنين ممن قاموا باستخدام هذه المعايير لتقويم أداء مجموعة من الطلاب.