0868

Not complete copy - just the basics

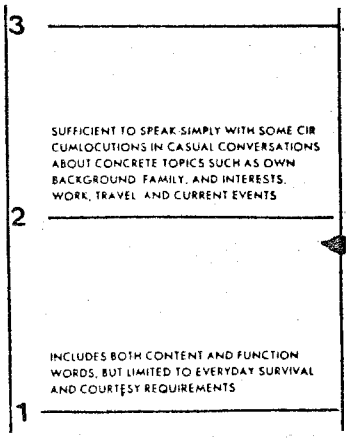
ETS ORAL PROFICIENCY TESTING MANUAL

Educational Testing Service
Princeton, New Jersey

(2) Language School Performance Profile

This document describes the characteristics of each linguistic factor by level and is particularly helpful in rating candidates whose language is stronger in one area than another, e.g., very fluent but poor pronunciation. The Performance Profile assists testers in looking at all of the components of a speech performance in combination with one other, and in not being unduly influenced by an area of particular strength or weakness. The Performance Profile is also a useful document for reporting test results to candidates.

To use the Profile, testers should read the descriptions for each factor and decide: (1) which description best fits the examinee's speech, and (2) where within the description the examinee's speech falls. For example, if a candidate's vocabulary fits the description for Level 1 but is fairly close to the description in Level 2, the tester would mark the Performance Profile in this way.



By doing the same for all of the factors, the tester can get an almost visual sense of the candidate's rating. Profiles of classic types of speakers at various levels are attached as Appendix VII.

(3) Grammar Grids

The grammar grids consist of language-specific tables of structures that are usually controlled by candidates at each level. "Grammar" here is understood in the widest possible sense. The elements that appear in the grids are those structures that experience proves to be consistently indicative of a given level.

Grammar grids have a number of important uses:

1. They objectify what is in the rater's mind when he/she rates an oral interview.
2. They facilitate training of new raters by letting them focus on specific problems when a general discussion of the global score is insufficient.
3. They are a tool for discussing why a particular rating has been assigned.

A number of disadvantages to grammar grids exist as well, however. The main danger in using grammar grids or other language-specific level-by-level guidelines is that one begins to conceive of the oral interview as a discrete-point test. Furthermore, it is impossible to capture all the necessary grammar structures in one grid or to indicate their relationship to each other. For this reason the grammar grids should be considered as only auxiliary aids in rating. Comparing the candidate's speech sample to the definitions is still and always will be the principal method of rating.

The starting point of a grammar grid is the government definitions, followed by the Guidelines for Assigning Language Proficiency Levels (pp. 33ff). The language-general descriptions provide the framework for language-specific statements. For example, the Level 1 definition

assumes that candidates must control present time; the Level 2 definition assumes that they must have ways of dealing with past and future time as well. It is the job of a grid, however to elucidate how time will be expressed in a given language. The grids have also been developed to respond to some questions that the language-general definitions and the Guidelines cannot address. For example:

1. At what point should Russian verbs of motion be mastered?
2. At what point should the distinction between preterite and imperfect in Spanish be under control?
3. At what point should humble and polite forms in Japanese be learned? At what level does one need both forms?
4. At what point should a German speaker properly use not only Sie and Du forms, but such titles of address as "Fräulein Helga, Sie...?"
5. At what point should a speaker control both the conjunct and disjunct pronouns of Italian and their ordering relative to one another?
6. At what point should word order be mastered in Mandarin?
7. At what point should the continuous tenses be controlled in English?

The grids also contain statements about extent of control of different kinds of grammatical structures. Each structure may be described from the point of view of full control, partial control, or concept control.

1. Full control. There are grammatical structures that must be controlled fully at a given level in order for the candidate to be a solid speaker at that level. An obvious example is

the regular present tense in a language like Spanish. A high 1 (Intermediate Mid) should use this structure with hardly any errors.

2. Partial control. Similarly, there are grammatical structures which may be only partially controlled at a given level. An example from French is the formation of irregular present tense forms like veux/veulent in which one still expects some errors at Level 1. Although it is perhaps hard to state in percentages, a grid should give some indication of the extent of control, often through the use of qualifying adjectives: some, most, etc.

3. Concept control. In most languages, there are a few structures that a candidate meets early in using the language, but may not control in all their details and subtleties until much later. In the Slavic languages, this may be aspect; in English, the continuous tenses; in Romance languages, aspect in the past tenses. For example, by Level 1+ (Intermediate High) in Spanish a candidate may show awareness of the concept of preterite versus imperfect but still fail to have all the forms correct in either tense. At Level 2 (Advanced) the candidate may have correct forms, but still much incorrect usage. Full control, particularly of subtleties, may come only at still higher levels.

Thus, one would expect the distinction of preterite and imperfect to appear in the Spanish Grammar Grid at several different points with a statement about controlling the concept, then some partial control of the distinction, and finally full control including subtleties.

It should be noted that not all grammatical structures will be commented on in positive form. In some instances, lack of control or a certain number of errors in a given structure is as indicative of a specific level as positive control. Thus, an Italian Grammar Grid could state that a Level 1 (Intermediate Level) speaker will probably be mixing genders of nouns and adjectives or matching singular verbs with plural subjects, whereas a Level 2 (Advanced) speaker would be expected to have mostly overcome these problems; and a Level 3 (Superior) speaker should control them almost all of the time (one or two errors in a twenty-minute test).

In this context, it is important to stress again that a single error counts nothing. It is only when such an error has been repeated consistently or fits into a larger pattern of errors that it should affect the overall score. Similarly, a single grammatical construction which is lacking or poorly controlled counts for very little. In fact, there is only one language, Spanish, where the lack of a single structure, the subjunctive, would keep the candidate from obtaining a Level 3 rating. This is true for Spanish because the subjunctive is used for a large number of Level 3 constructions expressed by other constructions in other languages. In the languages that the government has worked with so far, this is the only such case discovered. Used as an aid to the assignment of overall global scores, a grid can lead to more accurate rating and a better understanding of the overall oral interview system.

Conversation and Discrete-Point Orientation

To those accustomed to discrete-point testing, one may say that the oral interview is an integrative test, i.e., it addresses a number of skills

simultaneously and looks at them from a global perspective, rather than from the point of view of the presence or absence, control or semi-control of any given linguistic point. It is not that the linguistic points are ignored, but rather that they are viewed from the wider perspective of function.

Although it may appear, for instance, that testers set out to elicit an imperfect subjunctive form in Spanish, they are merely putting the candidate in situations where educated native speakers automatically use such forms. Then the candidate's performance is compared to that of an educated native speaker. Critical here is the fact that the oral interview is a test of usage, not of knowledge. We are rarely interested in whether candidates know how to form the imperfect subjunctive in Spanish or whether they can describe the contexts in which it might be used. Testers care only that the candidate use it at those times and in those contexts where educated native speakers do. Asking a candidate how to form the imperfect subjunctive, how to translate sentences that contain a particular construction, or how to complete a sentence where the completion should contain the desired form are not tests of usage. All are stages on the way, but the performance elicited has not been integrated into functional language situations. It is integrated usage that the oral interview strives to test and rate.

Grammar grids and other language-specific statements may seduce testers from the discrete-point tradition into preserving a discrete-point mind-set or applying it where unsuitable. The grids do, however, derive from experience in testing according to the language-general documents. The structures enumerated in grammar grids are the result of testing for a given function,

not for the discrete language features listed. In most languages, a constellation of factors (or various shifting constellations of factors) is what determines the rating.

Thus, to be a Level 3 in German, for example, one needs some combination of politeness subjunctives, hypothetical subjunctives, man constructions, passives, dependent clause word order, etc. To attain a Level 3 most of these have to be present, but a given test, while containing almost all of the above, might contain no man constructions at all. A conscientious tester from the discrete-point persuasion might specifically try to elicit such forms. But the definitions do not demand the presence of man constructions. The Level 3 rating is assured without it, and the time might be better spent in probing for still higher levels. Put another way, content validity in the oral interview rests not with testing for each of the points cited in the language-specific grammar grid, but in posing those topics or in placing candidates in those situations where educated native speakers use the structures in question.

There is a point when testers should consider discrete-point items. Two-thirds of the way through the interview the tester should do a mental review of the interview up to that point, summing up the candidate's pattern of strengths and weaknesses. As a result, the tester may be aware that a German speaker uses the man construction well, but rarely produces a passive; that the candidate readily employs politeness subjunctives, but uses the hypothetical subjunctives incorrectly. Furthermore, the candidate may seem to use rather simple vocabulary. The review may indicate that further questioning that might elicit subjunctives and passives/man constructions is not necessary, for the tester already knows about the strengths and

weaknesses in these areas. However, higher level vocabulary seems to be an area which could be extended by probing, posing higher level topics, etc., and might well be an area of exploration in the part of the interview before the Wind-Down. Again, this would not be approached by translations or asking the candidate to list ten high level words, but by posing those topics and situations which could be expected to elicit higher level vocabulary from an educated native speaker.

(4) Level 3 Descriptions

The Level 3 descriptions provide baseline information for higher level tests by giving language-specific profiles of Level 3 speakers. For the ACTFL/ETS system, in which candidates are designated "Superior" for all levels above 2+ (Advanced Plus), the Level 3 descriptions serve as top-of-the-scale criteria.

LEVEL 3 DESCRIPTION

FRENCH

Level 3 speakers should show an ease of delivery quite close to the one they have in daily dealings with people in their own native language, and their speech should not show any pattern of errors in high-frequency language structures such as the distinction between the passé composé and the imparfait, which should be correct most of the time. The following structures should also be correct most of the time: word order; pronouns (to him, to her, etc., whom, who, whose, etc., mine, etc., this, that, which one? what? etc.); prepositions; agreement (number and gender); and the use of c'est, il est, elle est.

A Level 3 speaker should have a vocabulary adequate to express and elaborate on his or her opinions. Speakers should have some knowledge of idiomatic constructions and colloquialisms. They should be able to paraphrase their ideas on unfamiliar topics without too much hesitation.

Note: For the level at which each of these features should be controlled, see the French Grammar Grid.

LEVEL 3 DESCRIPTION

GERMAN

Level 3 speakers should show an ease of delivery quite close to the one they have in daily dealings with people in their own native language, and their speech should not show any pattern of errors in essential grammatical features of modern standard German such as: gender, case endings, adjective endings, and agreement of nouns, pronouns, adjectives, and verbs; the present future, simple past, and compound past tenses of regular, irregular, and modal verbs; formal and informal speech, commands, and common subjunctive verb forms such as wurde, konnte, etc.; and all but the most complicated word order patterns.

They should also show at least some control of passive, man, and lassen constructions, and of both the hypothetical and indirect discourse subjunctive.

Level 3 speakers of German should have a breadth and precision of vocabulary adequate to express and elaborate on their opinions. They should have some knowledge of idiomatic constructions and colloquialisms. When discussing unfamiliar topics, they should be able to paraphrase their ideas without much hesitation.

Note: For level at which each of these features should be controlled, see the German Grammar Grid.

LEVEL 3 DESCRIPTION

ITALIAN

Level 3 speakers should show an ease of delivery quite close to the one they have in daily dealings with people in their own native language, and

their speech should not show any pattern of errors in high-frequency constructions such as: the present, future, perfect, and imperfect tenses; imperative forms; modal verbs; frequently used irregular verbs; passive and progressive forms; proper use of prepositions; agreements; word order; negative and interrogative forms; conjunctive and disjunctive object pronouns; and pronoun substitution.

Candidates must have fair control of low-frequency structures, irregular patterns and more sophisticated usage such as expressing opinion and feeling and the use of the conditional and subjunctive moods.

Level 3 speakers should also have adequate control of the sequence of tenses, enabling them to specify accurately the time of a given action in chronological perspective.

Level 3 speakers should have a vocabulary adequate to express and elaborate on their opinions. They should have some knowledge of idiomatic constructions and colloquialisms. They should be able to paraphrase their ideas on unfamiliar topics without much hesitation.

LEVEL 3 DESCRIPTION

SPANISH

Level 3 speakers should show an ease of delivery quite close to the one they have in daily dealings with people in their own native language. Their speech should not show any pattern of errors in high-frequency language structures, which should be correct almost all of the time: agreement of article, noun, and adjectives; the proper sequence of tenses ("He says he will come" versus "He said he would come"); the proper use of the distinction between the preterite and the imperfect; and the distinction in usage between ser and estar.

Moreover, the speaker should have a good command of most prepositions and be able to use the present subjunctive correctly most of the time and the imperfect subjunctive correctly about half of the time.

Level 3 speakers should have a vocabulary adequate to express and elaborate on their opinions. They should have some knowledge of idiomatic constructions and colloquialisms. They should be able to paraphrase their ideas on unfamiliar topics without much hesitation.

Note: For the level at which each of these features should be controlled, see the Spanish Grammar Grid.

CHAPTER 3: ELICITATION TECHNIQUE: THE ART OF INTERVIEWING

Elicitation, defined as the art of obtaining an adequate ratable sample from the candidate, most often takes the form in an interview of the tester asking a question or making a statement to which the candidate responds. The exchange should be friendly, and the interview should not sound like an interrogation (a danger at the lowest levels). On the other hand, while being as natural and relaxed as possible, the test is not simply a conversation. Testers must keep in mind at all times that their goal is to guide candidates to perform the functions indicative of their speaking level. If testers must choose between relaxed naturalness on the one hand and firm guiding on the other to get the necessary information, they should select the latter.

Each question in an interview has a purpose, and in fact often has more than one purpose. A question can encourage speech production by the candidate, can set the linguistic level, delineate a topic, elicit a speech sample in general or elicit in particular a specific function, a grammar point, or a vocabulary item.

Purposeless questioning leads to an unratable sample, no matter how much speech the candidate may produce in response. If each question posed by the tester has a purpose, then the chances of obtaining an adequate ratable sample and shortening the interview are greatly increased. Some questions, of course, may be wide of the mark. But purposeful, structured questioning improves the sample and sharpens the tester's feeling for the candidate's level.

The oral interview has been described as a single test with an infinite number of parallel forms. This is because the topics discussed in each interview can vary widely. Once an interview begins, no one knows how it will

develop. An interviewer must be prepared to pursue any topic or to go in any direction. Since testers do not rate factual content but only linguistic expression in any given area, they need not be subject matter experts to interview effectively. What is necessary, however, is to be mentally flexible and quick enough to pursue a topic in breadth or in depth if necessary, seeking out related topics, implications, and associated issues.

Once a topic is introduced a tester should stay with it, not moving on to a different subject until a ratable sample has been obtained. For example, an interviewer can sharpen a question's focus, raise its level, change its style, present provocative alternatives, ask the candidate's opinion--all on the same topic. Leaving a topic too soon can result in unconnected, discrete questions which resemble an interrogation more than an interview and which, in addition, will not yield a ratable sample.

At the lower levels, certain topics are mandatory. For candidates who are not able to sustain a conversation, testers should always turn to the "0+ Subject Areas." A survival or courtesy situation is always mandatory for Level 1 (Intermediate Level) tests. By Level 2 (Advanced Level) it is impossible to list all of the topics that might be introduced into the conversation.

Testers should begin an interview by posing questions in a normal tone of voice at the normal pace for the language. In lower level tests, it may be necessary to slow down, repeat, or paraphrase. Testers should resort to these adjustments only when a normal rate of speech clearly interferes with communication. Sometimes slowing down may actually hinder communication rather than aid it, since the candidate's short-term memory may not be able to retain as much material when it is spoken slowly. Testers should check their rate of

speech periodically throughout an interview to correct the natural tendency to speak more slowly or simply than is necessary.'

Once the candidate responds, the interviewer performs a number of steps nearly simultaneously. First, the tester must evaluate (assign a rating to) the candidate's response. This rating is provisional and becomes part of a kind of "running average." Particularly, testers should rate responses at the beginning of the interview provisionally, since early errors may disappear as the candidate warms up. Testers should ignore initial errors if the subsequent speech sample does not contain further examples of similar breakdown. Once the interviewer is aware of the level at which the candidate responded, he or she can evaluate the effectiveness of the question.

The tester should then ask himself or herself how well the candidate's response matched the purpose of the question.

A. Did the question elicit useful information about the candidate's language, i.e., a ratable sample? If so, why? If not, why not?

B. Did the candidate respond at the level desired? If so, why? If not, why not?

C. Have I exhausted what I need to know about the given function, grammatical point, vocabulary item, idiom, etc.?

D. Was the candidate interested in the topic?

The above steps place the interviewer in the position to formulate a new question, thus beginning the process anew and building on a series of impressions to formulate a definitive rating.

STRUCTURE OF THE INTERVIEW

Every oral interview follows the same general structure. This general structure guides the interviewer by directing his or her attention to certain

mandatory aspects of the test. An interview may be divided into four phases: Warm-Up, Level Check, Probes, and Wind-Down. The Level Check and the Probes take more time than the Warm-Up and the Wind-Down. At the very lowest levels, the limitations of the candidate's language may be such that the four phases will be indistinguishable from each other. At the very highest levels, neither Warm-Up nor Wind-Down will be necessary unless the candidate has not been speaking the language recently.

The Warm-Up. The Warm-Up consists of social amenities and simple conversation at a level that is very easy for the candidate. (At the lowest levels this may not be possible.) There are three purposes to this phase of the interview: (1) putting the candidate at ease; (2) reacquainting the candidate with the language if necessary; and (3) giving the interviewer a preliminary indication of the candidate's level.

For candidates, the main purpose of the Warm-Up is to put them at ease with the testing situation and to reintroduce them to the language. The length of the Warm-Up will depend on the circumstances; candidates who have not spoken the language for some time may need to get back into it gradually, while others may themselves immediately shift the conversation to a higher level. Testers should never skip this phase, but they may shorten it considerably if the candidate does not seem to need it.

One good way to begin the Warm-Up is for the tester to introduce himself or herself to the candidate in the target language. Since introductions are usually learned early in foreign language classes, it is easy for most candidates to respond, opening the way for further conversational exchanges.

For the interviewer, the Warm-Up serves the important function of giving a preliminary indication of the candidate's level. This preliminary indication

must be confirmed, because many candidates answer questions at the level and in the style in which they are asked. The best approach is for a tester to assume that the preliminary indication is to be checked in the next phase, the Level Check. In fact, the rest of the interview will be devoted to ascertaining whether or not this preliminary indication is accurate.

The Level Check. The purpose of this phase is to find the highest level at which the candidate can sustain a speaking performance. To find the level, the interviewer must test the breadth and depth of the candidate's ability in the language. How fluent is the candidate? How well does he or she pronounce the language? How accurate is the grammar? How wide is the vocabulary? How correct is the syntax? How native is the expression of ideas and concepts in the language?

Sometimes the level indication given by the Warm-Up is misleading, and the interviewer can begin the Level Check too low or too high. If the test begins at too low a level, the interviewer can simply raise the level of the questions and begin the Level Check over again. If the test begins at too high a level, the interviewer must bring the level down. Starting at too high a level is to be avoided, since bringing the level of an interview down is difficult to do without giving the candidate a sense of failure.

In the Level Check, testers should check a number of topics (both interest and non-interest areas) to see if the candidate can perform consistently at the level in question. Can the candidate accomplish the functions with suitable content and accuracy? When the candidate successfully passes the Level Check, his or her performance provides a floor to the rating. The next phase aims at finding the ceiling.

The Probes. The purpose of this phase is to make sure that the level the interviewer has been checking is the candidate's highest sustained level. To probe, the tester should take the candidate above the previous level several times in different ways: an involved question, a situation, a conversation between two testers into which the candidate is then drawn, etc. The interviewer may also want the candidate to ask some questions. If this phase has been successful, every candidate should leave the testing room feeling that he or she has been tested to the limit of his or her ability.

This phase is purposely in the plural because there should be several probes, at least three or four. Probes should furnish clear examples of linguistic breakdown. Sometimes the candidate actually tells the interviewer that the limit has been reached by saying, "I don't know how to say that in your language," or "I know what I want to say but I can't say it." In other cases, a sharp drop in fluency, a sudden groping for words, or a dramatic increase in grammatical errors give evidence of the linguistic breakdown.

If the interviewer has carried out the Level Check at too low a level, the candidate will probably be able to respond to the Probes consistently well. If this happens, then the interviewer must begin the process of Level Check and Probes over again and continue until the ceiling of the candidate's proficiency is found.

While the Level Check gives evidence of what candidates can do, the Probes show what candidates cannot do. Without this phase of the interview, candidates may appear to be more proficient than they really are. The Probes allow a tester to explain why a candidate's speech is not at a higher level, providing diagnostic information with specific examples.

Experienced testers learn how to interweave the Level Check and the Probes, so that the candidate is allowed to return to a level where

performance can be sustained before being asked another higher level question.

The Wind-Down. The purpose of this phase is to leave candidates with a feeling of accomplishment after stretching their speaking ability to the limit. It is also the tester's last chance to check out any aspect of the candidate's speaking ability that may still be unclear. Normally, the Wind-Down should return to the highest level that the candidate was able to sustain during the interview. It may even be helpful, particularly at the lowest levels, to end the test by returning briefly to a topic discussed previously. It is, of course, always appropriate to close by thanking candidates for the interview.

THE PSYCHOLOGICAL PLANE

The psychological plane refers to those aspects of the interview that encourage the candidate to talk. Once the Warm-Up is over, testers may not need to devote attention to this plane. If the conversation lags, however, it may be necessary to refocus on the candidate's level of comfort or discomfort.

Generally, candidates will be more inclined to talk when they feel comfortable psychologically. Thus, how candidates feel about their speaking ability and how they feel about oral testing are only two factors that may affect the interview. Among others is the tester's psychological stance with respect to the candidate.

Many interviewers are also teachers. The skills required by the two activities, teaching and testing, are not necessarily the same. Particularly, how the tester approaches the candidate in the test is different. Testers who are also teachers should develop the habit of changing their mind-set before beginning an oral interview. The helpful teacher who corrects candidates and

finishes their sentences for them will not be able to conduct effective interviews. A basic quality of friendliness is common to both good teaching and good testing. But testing demands an objective attitude on the part of the interviewer, an attitude which requires the candidate to prove that he or she can function independently in the target language.

Testers may need to remind candidates that they should respond in complete sentences or should speak a little louder. However, this should be the limit of coaching. Testers should also avoid giving clues. The candidate may be hunting for a particular word or form. By giving unnecessary hints, testers may fail to collect the information which will solidify the final rating.

Testers who are teachers may also be prone to correct candidates in the interview. Correction is unnecessary because it wastes valuable time and because the interviewers have already obtained a ratable sample--they know that the word or construction is wrong. Providing the candidate with a word or phrase should be undertaken only when the tester intends for the candidate to stay with a given topic and the word or phrase in question is necessary to continue the conversation.

Sometimes a candidate gets hung up on a word and asks for help. While testers may feel that they have to provide it, they run the risk that the candidate will repeat the request and thus impede the interview. A good strategy in this situation is to pretend to be a monolingual speaker of the target language and say, "I don't understand that. Can you explain what you mean another way?" This strategy has the double advantage of (1) not giving the candidate an easy escape when he or she doesn't know a word, and (2) encouraging the candidate to circumlocute, thus increasing the amount of ratable material.

One problem that testers who are also teachers often have is the tendency to fill pauses. This is a natural tendency in teachers who have learned to encourage communication by providing needed words or filling in when a student hesitates. In the oral interview, however, pauses may be ratable. A pause can signify either that the candidate has reached a point of breakdown, or that the candidate is thinking of the response. If the pause signifies that the candidate has run out of language, it is important, painful though it may be for tester and candidate alike, to allow enough silence to be sure that the candidate cannot manage the response on his or her own. This is particularly important when the interview is being taped for subsequent rating. It must be clear to the rater that the candidate was unable to cope with the question or topic.

In the case when pauses are productive, it is also important to allow the candidate time to think and to formulate a response. A tester who rushes to fill in the response may destroy ratable material. By allowing the pause to exist, the tester may garner a ratable sample.

A second type of tester intrusion is the interruption. Cutting off the candidate's answer also deprives testers of ratable material. Occasional short comments are permissible and might most suitably take the form of a minimal encourager (see below), but interviewers should constantly monitor themselves to make sure that they are allowing candidates to say as much as they wish in response to a particular question. Sometimes testers fall into a pattern of expecting only one-sentence answers, and thereby cut off the possibility of higher level responses.

A third kind of intrusion is the expression of an opinion on the part of the tester. This is allowable occasionally, and is particularly

useful in higher level tests in which the interviewer offers an opposing point of view in order to get the candidate to support an opinion. If overdone, however, tester opinions reduce the amount of speech by the candidate. Testers should check themselves during an interview to be sure that the candidate is doing most of the talking.

Another facet of the psychological plane is the way in which the candidate and interviewer interact. This interaction is a crucial aspect of interviewing, since some tester approaches can increase candidate discomfort and decrease the candidate's desire to talk. Below are a number of pointers to keep in mind during an interview.

Warm-Up

(1) Make the candidate feel at ease. Don't immediately launch into the question-and-answer format without some friendly words of introduction.

(2) Suggest some positive soothing action when the candidate appears nervous: "Would you like to move your chair, make yourself more comfortable?" Don't ask the examinee, "Are you nervous?" or say, "Gee, you really look nervous."

(3) Talk first about the weather, summer vacation, etc. Don't start off with a difficult question involving hard or obscure grammar, idioms, or vocabulary.

(4) Be warm, friendly, but objective. Don't be overly casual with the candidate if he or she is known to you either as a student or a friend.

(5) Act professionally about the candidate's previous language learning experiences. Never disparage the examinee's previous training, instructors, or testers.

Level Check and Probes

(6) Let the candidate talk--the candidate is the one who is being tested. Don't interrupt to interject your own thoughts or tell your own experiences. All of your comments should be calculated to increase or direct the candidate's speech.

(7) Act interested in the candidate and his or her experiences. Maintain eye contact; if you look toward the floor, window or the clock, the candidate will feel that you are not interested and will be less inclined to talk.

(8) Judge candidates only on the language in which they express their thoughts. You should not disagree with their ideas except insofar as your disagreement is calculated to encourage the candidate to express his or her ideas more extensively or to support an opinion. You should never play the role of authority, e.g., "I don't think you understood the Ukrainian culture, the truth is..."

(9) Follow up every clue that might lead to an area of interest. Probe for this as much as for levels. At lower levels, however, don't insist on a topic which is not the candidate's field. The candidate may become hostile and continue the rest of the test by answering in monosyllables. When this happens, the candidate may well be given a lower rating than he or she deserves.

(10) Find out what the candidate can do. Don't inhibit the candidate by correcting his or her grammar during the test.

Along with the interactions described above, there are some attentive behaviors, called minimal encouragers,[4] that interviewers can use to further communication simply and easily. They show the candidate that the tester is listening while being minimally disruptive and non-evaluative. They encourage the candidate to continue talking even though he or she may feel uncomfortable about the topic.

The use of minimal encouragers comes naturally if you focus on what the candidate is really saying. Recognize that, although you probably have all kinds of good things to say, it is most helpful for you as a tester just to be an effective listener.

Minimal encouragers come in two forms, nonverbal and verbal.

Nonverbal encouragers have the advantage of not interrupting speech or interfering when tests are recorded for subsequent rating. Some examples are:

(1) Keep eye contact, but avoid staring.

(2) Be alert. Exhibit an attentive (but not overly formal) body posture by facing and leaning slightly toward the candidate.

(3) Avoid looking at the clock.

(4) Nod.

(5) Smile.

Verbal encouragers are short words or phrases. They should be held to a minimum since they can be disruptive to a rater working with a taped interview. Some examples are:

(1) "M-mhuh."

---

[4]Irey, Allen E. and Jeremy Anthier. Microcounseling, 2nd edition. Springfield, IL, 1978. Pg. 78.

(2) "OK."

(3) "Good," "Fine."

(4) "Oh really?"

(5) "Go on/continue."

(6) Repeating one or two words.

(7) Rephrasing.

(8) Asking for clarification.

The words used should be non-evaluative, since evaluative statements are more intrusive and may cause candidates to stop and think or revise statements in the light of what they perceive the tester's evaluation to be.

Lists of non-evaluative statements similar to the above for English should be made for each language tested. Finally, minimal encouragers should be used to express sincere interest. Testers should take care to avoid sounding stiff or artificial.

THE ROLE OF CULTURE IN THE INTERVIEW

In the same way that there are linguistic thresholds that distinguish one level from another, there are also thresholds of cultural ability as one moves up the scale. In situations in government and education where the interview is used, it is most often, although not always, the case that the candidate is an American who has learned the target language. The higher candidates are on the scale, the more idioms they will know, the more culturally appropriate the vocabulary should become, and the more grammatically like a native speaker they should express themselves. Their allusions will become more culturally appropriate; they will express more nuances and subtleties appropriate to the culture; and their quotations, facial expressions and gestures will become more like those of a native speaker.

Opening topics (Warm-Up) are usually culturally American in focus. As the tester moves up the scale, he or she makes the gradual shift into target language culture. In the Warm-Up, the conversational topics are typically American, such as weather and family. Aspects of target language culture are purposely downplayed; for example, testers are objectively friendly, not effusive, as some cultures might demand. Target language questions common on first meeting in some cultures, but offensive or odd to Americans, are avoided, e.g., testers do not ask how much a candidate's house cost or how large his or her salary is.

For languages relatively close to English, such as Spanish and French, there seems to be a major shift in linguistic expression from American culture to target language culture at the 2+/3 border. For languages more distant from English, such as Japanese or Thai, this shift comes earlier, at about the 1+/2 border. At the crucial point on the scale, candidates should begin to react linguistically in terms of vocabulary choice, idiomatic expressions, structure, etc., more like a native speaker would react. The higher the candidate's speech on the scale, the closer it should approach that of an educated native speaker.

Culture affects all of the participants in an oral interview. Testers from a particular culture may feel constrained to pose questions in certain ways in order to be appropriately polite or elegant according to their culture. French testers, for example, often tend to make long, elaborate transitions from one topic to another. On the other hand, Russian speakers of other languages at lower levels tend to give answers which are too short--one investigation has shown that the majority of Russian answers to questions tend to be three words or less. The effects of these cultural differences can be

myriad, but aware testers can usually learn to minimize or eliminate them.

CONCLUSION

This chapter considered general elicitation techniques and the structure within which elicitation is carried out. The next chapter will deal with the questions and question types used to elicit the level-specific functions found in the verbal descriptions.

A NOTE ABOUT THIS MANUAL

Much of the content of this manual is based on the Manual for LS Oral Interview Workshops by Pardee Lowe, Jr., Chief of Testing at the Language School of the Central Intelligence Agency. That publication was developed for language tester training in the U.S. Government under the aegis of a joint CIA Language School/Defense Language Institute oral interview training project.

This manual has been produced for nongovernment use by Judith E. Liskin-Gasparro of Educational Testing Service (ETS). The provisional ACTFL/ETS rating scale, which makes the government oral proficiency testing system more appropriate for academic use, was developed by ETS under a grant from the U.S. Department of Education. The scale is being refined and pilot-tested by the American Council on the Teaching of Foreign Languages, Inc. (ACTFL), also with U.S. Department of Education support.

Both the ETS and the Language School/DLI manuals have been substantially revised in the last year in response to the needs of their audiences and as a result of discussions and workshops that have included various government and nongovernment agencies. It is anticipated and hoped that this cooperation and collaboration in the area of language proficiency testing will continue.

Table of Contents

APPENDICES

CHAPTER 1: INTRODUCTION TO THE ORAL PROFICIENCY INTERVIEW

RATIONALE FOR THE ORAL INTERVIEW

The oral interview is a test of an individual's foreign language speaking ability. The interview consists of a face-to-face conversation with one or two trained testers for a period of 10 to 30 minutes. The resulting speech sample is then rated on a scale of 0 (for no practical ability to function in the language) to 5 (for ability equivalent to that of an educated native speaker) with pluses given for performance stronger than halfway to the next level. The ratings are properly thought of as ranges, rather than points on the scale, since the description of proficiency at each level is broad enough to include weaker and stronger performances over a significantly wide range. The ratings were originally developed through a needs analysis of the tasks in Foreign Service positions involving the use of spoken foreign language.

The interview is a test of functional language ability, not passive skills or knowledge about the language. Although the descriptions of each level were originally designed for State Department purposes, they are general enough to apply to the evaluation of functional foreign language ability in other contexts as well. Modifications have been made to the lower end of the scale by the American Council on the Teaching of Foreign Languages, Inc., (ACTFL) and Educational Testing Service (ETS) for use with students in secondary school and college foreign language programs.

Why is an oral interview needed to assess speaking ability? What does it do that other tests cannot do?

Let us consider the following illustrations:

The Smith-Ronaldson Pen Company must send a team of sales representatives to France to introduce their line of ball-point pens. The company has learned through past experience that their sales representatives will be more successful if they can socialize with their potential clients in French. Moreover, the firm is concerned that they know enough French to find their way to appointments, deal with receptionists, and carry out simple telephone conversations in French when necessary. The chief manager chooses Fred Ainsley, who claims to know some French, to lead the team. Ainsley promises to take a refresher course. How can the manager know how well Ainsley will actually perform once he is in France?

In a second case, a university is planning to establish an overseas campus for Americans learning Spanish. While the university's Spanish department has several native Spanish speakers on its staff, none is able to take the time to set up and run such a program. An American faculty member volunteers, but can the administration be certain that this individual's Spanish is equal to the task?

The need in both of these situations is to know each speaker's current foreign language proficiency. Fred Ainsley may have had straight A's in his college French courses, but those grades tell us virtually nothing about his present linguistic abilities. Ainsley's willingness to attend a commercial language school may be helpful, although his boss has no way of knowing whether the school does indeed teach functional language skills. The university Spanish department will face a somewhat different problem in choosing the direction for its program abroad. There are many American Spanish professors in U.S. universities who can speak eloquently about

Calderón's dramas, but who might have serious difficulties with practical language situations, such as reading a Spanish contract or negotiating the terms of a lease with kitchen privileges for students who are living with local families.

The traditional paper-and-pencil tests do not answer the assessment needs presented in these two situations. The question that must be asked in each case is: Can the individual speak the language well enough to get the job done? The answer to the question lies in a foreign language test that will provide an immediate index of the current language proficiency of the examinee. The oral interview is such a test. It judges the examinee's performance against criteria characteristic of certain basic life situations with which a speaker of that language must deal. It thus furnishes the candidate, and the consumer who will use the candidate's services, with a strong statement of how the individual will function abroad linguistically in everyday life.

HISTORY OF THE ORAL INTERVIEW

Foreign language testing has always reflected the goals and methods of foreign language instruction. Before World War II, most foreign language instruction in the United States concentrated on the development of the literacy skills--reading and writing. Modern language instruction, introduced in the early years of this century, was quite similar to instruction in the classical languages. Students learned the rules of grammar, did translation exercises, and read materials of increasing difficulty.

Foreign language tests of the time reflected these desired outcomes. There were questions that required that the student recognize grammatical rules and identify grammatical constructions and parts of speech by name.

Often there were long vocabulary lists that asked students to find the English translation of a word in the language or the foreign language equivalent of a given word in English. Reading comprehension was usually tested in a multiple-choice format, often with the questions in English. Some writing exercises, usually translation rather than free compositions, were also included as a matter of course.

Foreign language capability became a pressing national need during World War II, and it became painfully apparent that adults who had been trained in grammar, translation, and appreciation of literature were not prepared to interrogate prisoners, comprehend radio broadcasts, or converse with our Allies. Language training at the Army Language School in Monterey, California, in which native speakers taught a variety of modern languages to armed forces personnel, was to become the catalyst for subsequent widespread changes in academic as well as government language curricula. The emphasis was on the development of oral skills, and the goal was the ability to speak the language for practical use.

After World War II, continued interest in training for functional language proficiency was concentrated primarily in the government. The Language School of the Foreign Service Institute, for example, had the task of preparing foreign service officers both in understanding the culture of the countries to which they were destined and in the acquisition of skill in the host country's language.

When it came time to evaluate these students, the FSI staff was faced with a difficult task. The training program emphasized oral skills and practical communication in the language. An instrument was needed that would allow these future diplomatic officers to function in the foreign

language in situations such as those they would encounter on the job. A course grade of A or B is meaningless to an ambassador, a chief of mission, or an artillery officer. The tests in existence at that time measured only literacy skills; there were no valid and reliable measures of speaking ability. Although we know, for example, that there is a relatively high correlation between listening comprehension and speaking ability and between reading and writing, the relationship is only true for large numbers of students. Individual decisions cannot be made on the basis of these correlations alone because there are many people who simply do not fit the mold.

Faced with the need to evaluate the actual performance ability of their graduates and the additional need to provide labels that would be readily understood by non-specialists, the linguists at the FSI Language School developed a rating scale to describe speaking ability and an interview-based evaluation procedure for assigning ratings. The scale is printed at the beginning of this manual and is discussed in greater detail in subsequent chapters. Once the scale and the original descriptions of each level had been decided, the next task was the development of a procedure whereby examinees could be placed on the scale. This procedure is the oral proficiency interview.

Beginning in the 1950s, the interview with its related scale became the measure used by the Foreign Service to describe the language ability of both students undergoing training at the Foreign Service Institute and others out in the field. The people tested through this procedure were not only Foreign Service officers but also personnel of other federal agencies. The FSI scale was adopted by other major agencies, such as the Central Intelligence Agency and the Defense Language Institute (the former Army Language School). At the Defense Language Institute, where large numbers of military personnel

were trained in a variety of languages, the interview procedure itself was at times impractical as a regular testing device and was replaced by a series of multiple-choice measures. Nevertheless, the descriptions of ability used by DLI paralleled and were essentially the same as the Foreign Service Institute 0-5 scale.

The oral interview and the notion of language proficiency testing spread throughout the government because the measure and the scale were equally valid for those who went through the FSI program and those who learned the language at home, in school, or overseas. The interview and rating constituted a summative evaluation procedure independent of any program of studies. The criterion against which performance was to be compared was in every case the speaking ability expected of an educated native speaker of the language.

In the late 1960s, the Peace Corps turned to the FSI for its language assessment needs. The Peace Corps at that time was teaching languages to thousands of volunteers at training sites in the United States and abroad. The FSI provided personnel to test Peace Corps volunteers at various stages during their training and service: typically, at the beginning, the mid-point and the end of training and then after one and/or two years in the field. When the magnitude of the task strained government resources, ETS was asked to carry on these language testing activities. Senior ETS staff were trained by the FSI in the interviewing procedure and provided at first testing of trainees and volunteers for the Peace Corps, and later on the training of testers in the various Peace Corps languages for on-site testing by the Peace Corps. The Peace Corps was one of the first large-scale oral interview testing activities not operated directly by a government agency.

In the early 1970s the possibility of wider application of the oral interview procedure came from bilingual education programs and agencies at the municipal and state levels. ETS developed for the State of New Jersey an oral proficiency testing program in English as a Second Language and in a number of other languages as part of the certification procedure for bilingual and ESL teachers. A similar program was developed in Spanish for the Texas Education Agency. The states of Illinois and Massachusetts and a number of local and county school districts in California have used and continue to use the oral interview for bilingual teacher certification, hiring, or placement.

Other educational applications of the oral interview have been found by the Language Training Mission of the Church of Jesus Christ of Latter Day Saints, which contracted with ETS for a series of training sessions in a variety of languages for missionary work. The province of New Brunswick, Canada was the first state or provincial government agency to use the oral interview procedure for evaluating second-language skills of secondary school students. In addition, the Experiment in International Living and a number of other agencies in recent years have also requested and received training of their personnel for conducting oral proficiency interviews.

THE ORAL PROFICIENCY INTERVIEW FOR ACADEMIC USE

As the government scale, currently known as the ILR (Interagency Language Roundtable) scale, attracted interest within formal academic circles, a feeling grew among foreign language professionals in secondary and post-secondary education that the scale was not as sensitive at the lower end as it might be. The ILR, or government scale, covers the whole spectrum of speaking ability from Level 0 ("no functional ability") to

Level 5 ("ability equivalent to that of an educated native speaker of the language"). In the assessment of the speaking ability of high school and college students, the full range of the scale is rarely used. Most students, even those with such extracurricular experiences as a summer or a year abroad, rarely score above level 2 or 2+ on the ILR scale.

John B. Carroll, in a study reported in Foreign Language Annals (v.3, #2, December 1967), assessed the foreign language proficiency of undergraduate majors of French, German, Russian, and Spanish. Most were rated 2 or 2+ at the end of their senior year. Since most students of foreign languages are not majors nearing graduation from college, it seems safe to assume that the ILR levels of most interest to high school and college foreign language teachers will be levels 0 and 1.

The ILR scale as it is used in the government makes provision for five definable ranges of proficiency between levels 0 and 2: 0, 0+, 1, 1+, 2. These ranges are not equidistant from each other, but rather are farther and farther apart as one moves up the scale. Thus, 0 and 0+ are relatively close to each other, while the gap between 1+ and 2 is relatively great. Even 0 and 0+, the ranges closest to each other in the government system, are so far apart that they may not be sensitive to differences in oral ability among students. An informal study conducted by ETS, in which oral proficiency interviews were administered to approximately 30 high school Spanish students, confirmed the hypothesis that the lower end of the scale did not effectively discriminate among students whom teachers would judge to be significantly different in oral ability.

In recognition of the interest on the part of the academic community in the scale and the interview, the Foreign Service Institute held a series of

three Testing Kit Workshops in 1979-80 for college teachers of French and Spanish. The faculty members were introduced to the ILR system of oral proficiency assessment and were trained to interview and rate. Support was offered via mail and telephone to the professors when they returned to their campuses and began to test their own students.

One major result of both the ETS study mentioned above and the FSI workshops was consensus on the need to expand the lower end of the ILR scale to make it more applicable to students in traditional academic environments. There had to be more points between levels 0 and 2 so that students' progress over the course of a semester or a year, while perhaps not enough to be reflected in a new ILR rating, could still be registered. There was also some discussion about incorporating a mechanism to provide diagnostic information to students about areas of strength and weakness.

At about the same time, ETS was approaching the question of an academically oriented speaking scale from another perspective. In a project called "The Common Yardstick," the English Speaking Union of Great Britain, the British Council, Educational Testing Service, the Deutscher Volkshochschulerband and representatives of the U.S. government and various business and academic groups met to develop or adopt a series of descriptors of language ability. The United States, the British Council, and the English Speaking Union all presented draft scales for consideration by the group. Subsequent to two meetings in Great Britain and the United States, a grant to ETS from the U.S. Office of Education provided for further refinement of the Common Yardstick scales. The provisional ACTFL/ETS scale in academic use that appears at the beginning of this manual is the product of the Common Yardstick project.

In 1981, the ACTFL received two grants from the U.S. Department of Education in the area of foreign language proficiency. One of these, "Professional Development: Foreign Language Oral Proficiency Testing and Rating," has built on the work begun by the FSI Testing Kit Workshops and has undertaken the training of college foreign language teachers in oral proficiency testing. At a workshop for professors of French and Spanish in February 1982, the revised scale with the expanded lower end was taught for the first time.

The second ACTFL project, "A Design for Measuring and Communicating Foreign Language Proficiency," has created generic and language-specific goals for speaking, listening, reading, writing, and culture that can be used as a graduated sequence of learning steps in instructional programs. The ETS speaking descriptions developed during the Common Yardstick project were turned over to ACTFL for use in this project, and have served as the basis for the generic goals in speaking and as a model for the goals in the other skill areas.

The growing interest in oral proficiency testing is evident in the burgeoning number of workshops, conference sessions, and training sessions that have taken place in the past two years and that are planned for the near future. The ETS Division of Educational Services is sponsoring a series of two-day familiarization workshops in oral proficiency testing; all of the regional and national foreign language conferences have featured and plan to include more sessions and workshops in this area. A number of institutions and agencies, in most cases with support from ACTFL, are planning further work in the development of oral proficiency testing skill on the local level and in the exploration of the implications of the oral interview on curricula and methods of language teaching.

COMPARING THE ORAL INTERVIEW TO STANDARDIZED TESTS

It is possible to separate foreign language testing instruments into the two broad categories of achievement tests and proficiency tests, according to the particular kinds of information that the instruments and procedures are intended to provide.

The purpose of language achievement testing is to determine students' acquisition of various specified aspects of course content. Achievement tests can range in scope from short quizzes to chapter tests to final examinations covering the content of a whole course. Because the body of material in an achievement test is limited, it is possible to earn a perfect score. Those who have mastered the material to the same degree will receive the same score.

An achievement test is a test for which one can prepare. A strict achievement test never asks questions on material that has not been covered. The distinguishing characteristic of all achievement tests is that they are based on and reflect specified, predefined elements to which the student has been exposed in the course of the teaching process. Results on achievement tests are expressed as percentages of correct answers or scaled scores. Norm-referenced achievement tests compare students to one another. Criterion-referenced tests measure students' performance against the standard of mastery of course content.

The oral interview is not an achievement test, but a proficiency test. The purpose of language proficiency testing is to assess the examinee's language performance in terms of the extent to which he or she is able to use the language effectively and appropriately in real-life situations. In contrast to achievement testing, in which the test materials

are based on the content of a specified curriculum, proficiency testing is curriculum-free; it focuses exclusively on language competence without regard to the place, length of time, or manner in which that competence has been acquired.

Since a proficiency test does not cover any specified body of material, it is not possible to prepare for it. An oral proficiency test will test everything an individual knows about how to use the language by sampling his or her speech production on a variety of topics at a number of levels. An individual can get a "perfect score" on an oral proficiency test only by demonstrating speech production equivalent to that of an educated native speaker of the language.

In a proficiency test, candidates will always be asked questions for which they are not prepared. This is because the tester's role is to get a sample of the best language of which the candidate is capable. The probes into linguistic areas in which the candidate is not prepared find the limits past which he or she is unable to go.

RELIABILITY AND VALIDITY OF THE ORAL INTERVIEW

In the years following World War II, testing in the United States became more a science than an art. Multiple-choice tests grew in popularity because of their high reliability. Reliability in this context is understood to mean score stability in a test/retest situation. Examinations requiring free responses, such as written essays or taped speech samples, were rejected on scientific grounds as unreliable. Human beings simply do not measure the same written or spoken sample as consistently as a computer can scan multiple-choice answer sheets.

In developing or choosing a testing instrument, the demands of validity must be considered along with the need for reliability. The oral interview was originally developed because the more reliable paper-and-pencil tests were not valid measures of oral production. A valid instrument in this context is understood to mean one which actually tests what it purports to test. This is known as content validity. Face validity, also a characteristic of the oral interview, has to do with whether a test has the appearance of doing what it purports to do, or feels to the candidate like a valid test. A well-structured oral proficiency interview tests speaking ability in a real-life language context--a conversation. It is almost by definition a valid measure of speaking ability.

Content validity should also be thought of as stability or consistency of content from one test to another. The oral interview is not a fixed series of questions; the topics and the questions asked vary from one interview to another. Well-trained testers, however, will administer interviews that can be thought of as parallel forms of the same testing procedure. Although the questions and the topics may differ, and indeed should differ, from test to test, the question types remain the same. For example, the question, "What would you do if you won $10,000 in the lottery?" presents the topic "unexpected financial gains" and the question type "hypothetical question." Rather than repeat the topic and risk test compromise, a tester could use the same question type with a different topic: "If you were starting college now, what course of study would you undertake?" This strategy avoids test compromise while maintaining content validity.

If different interviewers were not looking for or observing the same linguistic behaviors, content validity would indeed be weakened. However,

workshops that train potential testers in the elicitation techniques and provide them with a common understanding of the standards will strengthen the content validity as well as the reliability of the interview.

Inter-rater reliability, a major concern in all tests of production, is the degree to which two testers listening to the same interview will assign it the same rating. Studies at ETS and in the government show that experienced testers may differ by a plus point in assigning ratings to the same interview. Extensive standardized training and periodic recalibration have successfully assured the reliability of both interviewers and raters.

CHAPTER 2: RATING ORAL PROFICIENCY INTERVIEWS

The oral interview is an integrative test, not a discrete-point measure. The ratings are expressed in global terms, by comparing the totality of a candidate's speaking performance to the descriptions at each level. Although the ratings are holistic, they do include awareness of the factors contributing to overall performance: pronunciation, fluency, grammar, and vocabulary. Still, the global rating is always primary in the evaluation of oral interviews. The factors must be considered in light of how they are integrated into the candidate's speech.

PRINCIPLES OF RATING

The level descriptions are not points on a scale--they are ranges. Since speech production is more than a collection of discrete utterances, it is not possible to make overly fine discriminations. Candidate A may be stronger than Candidate B, but both fulfill the requirements for the same level and fail to fulfill them for the next level.

Within an interview a candidate's performance usually varies, showing both areas of strength and areas of weakness. The tester's job in each interview is to find the highest sustained level at which the candidate can speak. This is the highest level at which the candidate sustains most of his or her speech sample during the test.

In a test candidates will demonstrate at moments that they can go beyond their highest sustained level. The peak level a candidate will occasionally reach is defined as the uppermost level at which a candidate may perform on an isolated topic or topics in the test. The tester must discover whether or not the candidate can sustain performance at this peak level, (i.e., whether it is

in fact the candidate's sustained level). This is done by means of the probes, which push candidates into areas where they may not have the language to sustain the highest level they have reached until that point in the test. When they clearly perform at a lower level for most of the test, then that sustained lower level is the rating.

Example: Marilyn Quirk discourses brilliantly on the differences in economic systems between France and the United States.

The tester rates her performance on this topic at Level 4. However, probes on other topics, such as the nature of French and American democracy, comparison of cultural reactions to women's liberation, and unknown situations such as exchanging a wrong size suitcase given as a present, show that the candidate sustains her performance at Level 3+. The final rating is thus 3+.

In the ILR system, plus levels are used when a candidate's performance substantially exceeds the requirements of a given level and when the candidate produces speech at the next higher level but does not sustain it at this higher level. Graphically expressed, a plus range is relatively narrow, i.e., it represents an area smaller than the upper half of the base level. For example, a 1+ would be represented by 1.6-1.9999.... The following examples of candidate performance illustrate the rule for assigning plus levels in the ILR system.

Example: At times during the test Peter Lee was a 2, but he frequently spoke in ways characteristic of Level 1. (He should be assigned a 1+.)

Example: Mary Smith is a 3 in vocabulary and fluency, but her grammar does not exceed 2. (She should be assigned a 2+.)

Example: John Hart is a 3 in grammar and fluency, but his vocabulary does not exceed 2. (He should then be assigned a 2+.)

In the ACTFL/ETS system, the plus levels have been incorporated into the definitions at the Novice and Intermediate Levels. Thus, Novice High is equivalent to the ILR 0+, and Intermediate High is equivalent to the ILR 1+. At the Advanced Level, the procedure for assigning the plus level follows the rule described above, and Advanced Plus is equivalent to the ILR 2+.

In evaluating a speech sample, raters look for patterns of strength and patterns of weakness. One example of strength or weakness will never determine a rating. During an interview, a candidate who shows a single strength should be asked questions that will demonstrate whether this strength is an isolated occurrence or whether it is part of a pattern of strengths that will affect the rating. Consider the following examples:

Example: Tom Hanson used one Spanish verb in the preterite tense correctly (yo fui), but his speech is otherwise consistently at the Intermediate Level (ILR Level 1). He should be rated Intermediate (Low or Mid, depending on the speech sample) or ILR Level 1. One strength counts nothing.

Example: Daisey Jones uses several complicated word order constructions in dependent clauses in German, along with passives and a man construction, but cannot sustain these Level 3 (Superior Level) constructions in what is otherwise a Level 2 (Advanced Level) test. This is a pattern of strengths, so she should be rated 2+ (Advanced Plus).

Similarly, one error counts nothing. A single error will not disqualify a candidate's speech sample from a higher level, but a pattern of weaknesses or errors will.

Example: Michael Benson shows a pattern of errors in French that is characterized by lack of noun-adjective and subject-verb agreement, and wrong forms and usage of the passé composé and imparfait. He demonstrates strength in pronunciation and vocabulary at Level 2 (Advanced Level). Since his weaknesses are at the 1+ Level (Intermediate High), and since they form a pattern, he should be rated 1+ (Intermediate High).

Since language learning is such an individualized process, there are innumerable combinations of linguistic features that make up ratings at each level. In addition, the way in which a language is learned will affect the constellation of linguistic behaviors, particularly at the Intermediate and Advanced Levels (Levels 1 and 2). Beyond the Advanced Level (Level 3 and above) a solid core of proficiency is expected, which will make for less variability in the characteristics of speakers at these levels.

A distinction is often made between candidates who learn the language in school and those who learn it by living in the target language environment. At the Intermediate Level (Level 1), for example, those who have learned the language in school will usually be relatively strong in grammar but lack everyday vocabulary. An individual who has learned the language by living in the target language culture, on the other hand, will usually know a lot more vocabulary but will use incorrect grammar.

By the Advanced Level (Level 2), the differences between these two types of learners may be even more pronounced. The individual who has learned the language in school still evidences the same relatively weak vocabulary and strong grammar. The person who has learned the language through language use

alone will at this level have a large everyday vocabulary and a good ability to communicate. The person's grammar, however, will have serious flaws, and often these flaws will have become permanent, or "fossilized." Such speakers often speak very fluently, although incorrectly, and are hard to rate because they seem more proficient than they really are.[1]

It is important in rating non-school learners to listen below the flow of speech as well as to it. Good pronunciation and fluency can mask weak grammar and sometimes vocabulary. Non-school learners tend to be high in the "integrative factors"--they are able to use the language they have effectively.

The principal procedure involved in all ratings is to compare the speech sample in the interview with the level descriptions to find the closest match. However, there is a tendency for testers to compare candidates at a given level to each other, rather than to compare each candidate with the descriptions. The danger with comparing one candidate to another as a basis for assigning ratings is that the standard will be subject to drift.

The parable that follows[2] illustrates how standards can drift when ratings are based on factors other than the level descriptions.

The Story of Ice Cream

In the land of Ice Cream there was once a law that stated that the designation "creamy ice cream" could be applied only to those batches of ice cream whose butterfat content was 30 percent or higher. Two ice cream testers were designated to check each batch so that the law could be applied uniformly.

---

[1]Graphic representations of the speaking performance of non-school learners are included as Appendix VII.

[2]The author of this parable is Pardee Lowe, Jr.

One day, however, an ice cream manufacturer said that a batch he had produced, which the testers showed had 27 percent butterfat content, was in fact a true "creamy" ice cream. He said, "Taste it for yourselves!"

So a group of four managers plus the testers inspected the ice cream. One pointed out that it was "creamy" to touch. Another indicated that it had a "creamy look." Two said that it "tasted creamy." The two testers said that the butterfat content was 27 percent, which did not conform to the standard for the designation "creamy." One of the testers did admit that such a batch of ice cream might "taste creamy."

But by this time the managers were feeling sorry for the ice cream manufacturer because he had tried so hard and had just been passed over for a promotion, and they declared that the batch of ice cream was indeed "creamy."

P.S. In subsequent years, the definition of "creamy" ice cream became so completely divorced from the butterfat content that the inhabitants of that land discovered one day that ice cream and ice milk were differentiated only by the packaging.

Two candidates can give quite different performances and still be at the same level. The most extreme cases are those of the school and non-school learners described above, but any two individuals will use the language at their disposal differently. Practically no speech sample will exactly match a level description in all respects. However, it is possible for two samples to be assigned the same rating if they show patterns of strength and weakness that fall within the boundaries of a given description, e.g., Candidate A is weak in vocabulary but strong in grammar and Candidate B is strong in vocabulary but weak in grammar. Two samples could also be

assigned the same level by being at different points within a given range, e.g., Candidate X is a weak Level 2 (Advanced Level) and Candidate Y is a strong Level 2 (Advanced Level), although neither one is strong enough to merit a plus level rating.

FUNCTIONAL TRISECTION

The study of the oral interview makes explicit for training and research purposes the factors that testers consider implicitly in assigning the global ratings. In assigning ratings, testers consider the candidates' overall speaking performance. They must also consider the influence of unevenness (peaks and valleys) on the overall impression and therefore on the rating.

What exactly constitutes the "overall speaking performance?" The oral interview tests a candidate's ability to function in specified contexts with suitable content and accuracy. Consequently, a candidate's linguistic behaviors may be viewed from three different vantage points: the functions, the context in which they occur (including the specific content), and the accuracy with which the functions are accomplished. For example, a candidate could be asked to talk about a trip he or she took (a Level 2/Advanced Level task). The FUNCTION is narration in the past; the CONTEXT is personal experience on a trip; and the ACCURACY acceptable at Level 2/Advanced Level is set forth in a verbal description against which the candidate's performance may be judged.

The verbal descriptions of each level contain statements about the function, content, and accuracy requirements for that level, but are not documented in detail. More recently, the CIA Language School has developed a "Functional Trisection of Oral Proficiency Levels" that is a useful companion to the government definitions.

FUNCTIONS

In the context of the oral interview, functions are equivalent to linguistic tasks, such as asking questions, giving information, describing, etc. Functions are non-job specific. Asking questions, for example, is a linguistic function that is common to educators, doctors, and engineers, although the context in which each of them would pose questions might well vary considerably.

Given a flawed linguistic performance in an interview, it is tempting, although possibly misleading, to point to grammar or pronunciation as the determining factor in a low rating. In most cases, the striking feature of a flawed performance is that either communication failed to take place, or that it was so incomplete that a partial misunderstanding resulted. The functions that the candidate undertook were not accomplished.

In ascending the scale, a distinction can be made between major borders between levels and minor borders within levels. The 0/Novice Low/Novice Mid (0/0+) boundaries, for example, are minor borders characterized by a gradual shift from no ability to use the target language to ability to function in it. Some utterances may remain unintelligible, while others become progressively more intelligible.

A major border, on the other hand, is characterized by a threshold, and as such there is a constellation of factors (a new interrelationship of function, context and accuracy) that marks the change from one level to another. At the Novice High/Intermediate Low (0+/1) border, for example, there is a clear delineation between memorized material (Novice or 0+) and creating (Intermediate or 1). An Intermediate Level (Level 1) speaker can create a large number of sentences with the same language that appears in a Novice

Level (0+ Level) memorized phrase. In addition, a candidate at the Intermediate Level (Level 1) provides the necessary grammar and vocabulary to communicate, rather than only the material, as a Novice Level (0+) speaker would.

The functions characteristic of each level will be treated in detail in the chapters devoted to level-specific material. Listed below is a summary of the functions characteristic of each level. Note that the functions are non-job-specific.

CONTEXT/CONTENT

Linguistic functions do not exist in isolation, but rather are always attached to a particular context or topic. The content of an oral interview is the most variable element, depending in large part on the interests and inclinations of the candidate. To determine whether a candidate could accomplish the Level 2 (Advanced Level) functions of narrating and describing, a tester might ask a bilingual teacher to talk about the steps she took to prepare herself for her current job. The tester might ask a high school or college student how he spent last year's summer vacation. In both cases, the function--narration and description in the past--is the same, but the content is different.

The oral interview is a test of general language ability. From Level 3 up, it is assumed that a candidate can make statements about any general area, that is, his or her language is sufficient to support an opinion, hypothesize, or at least state a disavowal of interest in discussing a particular topic. The interview does not rate the factual accuracy of statements, but rather the language usage revealed in discussing a given topic. It is important when rating to evaluate the content of a candidate's response in light of what the response would have been in the candidate's native language. For example, the question "If you were the President of the United States, what type of tax legislation would you propose, and why?" might receive little or no coherent response in any language from one candidate, while another who is interested in economics might have a ready response but not sufficient command of the language to express it.

At the lower levels, the oral interview may resemble an achievement test because of the closer connection between what candidates have learned

in academic courses and what they are able to talk about in an interview. The tester's task is to discover an area of interest to start a candidate talking and then to turn the conversation to a number of other areas to furnish an adequate sample of how the candidate functions in the language outside of the one or two favorite topic areas.

The content areas characteristic of each level will be treated in detail in the chapters devoted to level-specific material. Listed below is a summary of the content areas characteristic of each level.

ACCURACY

Accuracy refers to the acceptability, quality, and precision of the message conveyed. Grammar is a major component of accuracy, although not the only one. For example, an individual may write perfectly correct English, but in speaking may not pronounce the third person singular present tense -s or the past tense marker -ed. In these cases accuracy in oral production is affected by poor pronunciation, not poor grammar.

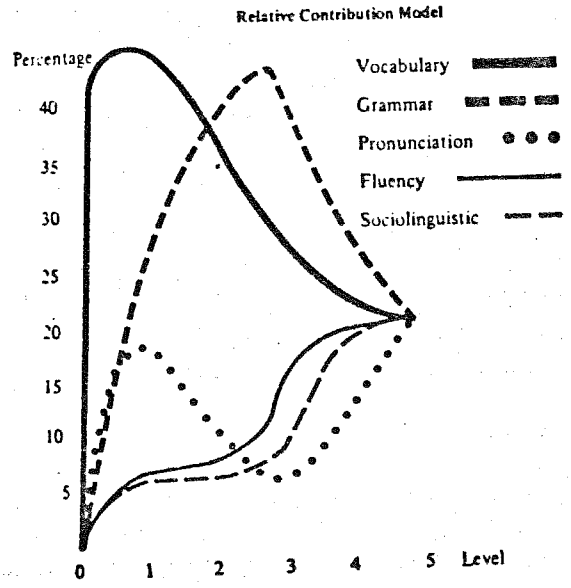
At the lowest levels, accuracy can be defined as intelligibility. At the Novice Level, virtually no utterances are grammatically accurate, and the native speaker who receives the message must rely on experience and context to interpret what is said. If the native speaker knows the errors a speaker from a particular language background is likely to make, then the native speaker will be more able to bridge the gap between the message conceived by the candidate and the message received.

By the Advanced Level (Level 2), accuracy becomes a matter of correctness, rather than intelligibility. Candidates at this level do not require an interlocutor accustomed to dealing with foreigners. The notion of communicating in spite of the inaccuracy of the message ceases altogether at the 2+/3 border.

The accuracy characteristic of each level will be treated in detail in the chapters devoted to level-specific material. Listed below is a summary of the accuracy statements for each level.

MODEL OF RELATIVE CONTRIBUTION OF LANGUAGE FACTORS

The linguistic factors that contribute to any speech performance are vocabulary, grammar, pronunciation, fluency, and the sociolinguistic factor. The relative importance of these factors varies as one ascends the rating scale. For example, at the very lowest levels, vocabulary, grammar, and pronunciation are most important for successful performance, while the other factors are relatively less important. At Level 5, the very top of the scale, all five factors or subskills will contribute equally to successful performance. The relationships among the five factors at all five levels are depicted in the figure below.[3]



(Note: This hypothesized model is most applicable to Indo-European languages.)

---

[3]The research that led to the development of the Relative Contribution Model was carried out by Ray Clifford and is reported in Higgs, Theodore V. and Ray Clifford, "The Push Toward Communication." In Higgs, Theodore V. (ed.), Curriculum, Competence, and the Foreign Language Teacher, v.13, ACTFL Foreign Language Education Series. Skokie, IL: National Textbook Company, 1982. Pages 57-79. The model depicted above is on p. 69. The discussion that follows is excerpted from the article.

In interpreting the figure above, it is important to remember that the height of a curve at any given proficiency level indicates relative contribution for each subskill. The fact that the vocabulary curve drops as it approaches Level 5 does not mean that less vocabulary is needed, but that in comparison to the other four contributory skills, vocabulary declines in relative importance. Because this is a graph of relative contributions, the values of the five curves at any given level always total 100 percent. At Level 1 (Intermediate), the most important component of the 100 percent is vocabulary, followed by sufficient grammar to create with the language, and a minimum threshold level of pronunciation sufficiently accurate to be understood. Fluency as measured in terms of words per minute and sociolinguistic elements are not yet crucial, because at this level one is concerned with listeners who are used to dealing with foreigners, and the expectations of both the speaker and the listener are quite low.

At Level 2 (Advanced), these relationships shift. The relative contribution of grammar increases as the required linguistic tasks (i.e., the range of linguistic functions to be mastered) become more complicated. At the same time, the relative importance of pronunciation begins to decline after reaching the minimal level required to be understood.

At Level 3 the relative mix of contributing subskills changes again. Grammar is more important than vocabulary, and the importance of the subskills of fluency and sociolinguistic sensitivity increase. Although one can still succeed with a foreign-sounding pronunciation, it is necessary to possess sufficient sociolinguistic skills and fluency so as not to offend or bore one's listeners.

At Level 4, the curves begin to coincide as functional performance approaches the level of the educated native speaker, who by definition controls each of these language aspects perfectly. On the assumption that a person whose language was lacking in any one of these areas would not be judged to be an educated native speaker, the investigators hypothesized that all subskills would contribute equally to the global performance rating of Level 5. Thus, all component curves converge at this level.

USEFUL DOCUMENTS FOR RATING

The government descriptions of oral proficiency began as one-line statements. These definitions were later expanded. Experience has indicated that additional guidelines for raters are often helpful, particularly with regard to quality or accuracy. In addition, the definitions are language-general, and certain language-specific statements provide important supplemental information.

Over a long period of time the government has developed several types of supplementary guidelines for raters that have been included below. These guidelines should be used only as supplements, not as replacements for the official level descriptions. They may also be useful in helping testers report diagnostic information to candidates after the interview.

(1) Guidelines for Assigning Language Proficiency Levels

(Speaking-Understanding)

Note that these guidelines refer to the ILR scale only. For raters working with the provisional ACTFL/ETS scale, they will be useful only as an aid in establishing the base level, not the finer distinctions. Note also that the guidelines include understanding (comprehension) as well as

speaking. The understanding guidelines will be useful for interviewers in deciding what questions to ask and at what speed.

LEVEL 0

SPEAKING: The examinee has no practical speaking proficiency. May have a few isolated words and phrases which are of no practical use. UNDERSTANDING: The examinee understands some isolated words and phrases, but is unable to participate even in a very simple conversation.

LEVEL 1

SPEAKING - Subject Matter: The examinee has the minimum proficiency for survival on a day-to-day basis in the target country, i.e., functions in simple question-and-answer situations. Knows enough at this level to satisfy ordinary courtesy requirements. Able to ask and answer questions relating to situations of simple daily life and routine travel abroad. The examinee is also able to handle requests for services such as renting a hotel room and ordering a simple meal. SPEAKING - Quality: The examinee at this level normally makes errors even in structures which are quite simple and common. Vocabulary is limited to the type of situations mentioned above, and even in these situations he or she sometimes uses the wrong word. Although pronunciation may be poor, he or she makes the minimum constrastive distinctions, including stress, intonation and tone patterns necessary to be understood. UNDERSTANDING: The examinee is able to understand simple questions and statements relating to simple transactions involved in situations of daily life and independent travel abroad, allowing for slowed speech with considerable repetition or paraphrasing.

LEVEL 2

SPEAKING - Subject Matter: The examinee is able to talk in some detail about concrete subjects such as own personal and educational background, family, travel experiences, recreational activities, and familiar places. SPEAKING - Quality: The examinee has enough control of the language to be able to join sentences in limited discourse. Good control of the morphology of the language (in inflected languages), and of the most frequently used syntactical structures. Although vocabulary is sufficient to talk with confidence about the type of topics described above, the limited vocabulary fairly often reduces the examinee to verbal groping, or to momentary silence. A foreign intonation and rhythm may still be dominant. UNDERSTANDING: The examinee is able to comprehend questions and statements relating to common social topics, when the language is spoken at normal conversational speed. Can get the gist of casual conversations with educated or well-informed native speakers talking about subjects on the level of current events, allowing for occasional repetitions or paraphrased statements.

LEVEL 3

SPEAKING - Subject Matter: The examinee is able to converse and express opinions about such topics as current events, including political and social problems of general nature. SPEAKING - Quality: The examinee has good control of grammar, though there are occasional errors in low-frequency structures and in the most complex frequent structures. The vocabulary is broad enough so that he or she rarely has to grope for words in discussing the topics mentioned above. A foreign phonology, though apparent, is no longer dominant. UNDERSTANDING: The examinee can comprehend most of what is said at a normal conversational rate of speech. A person at this level

is able to understand to a high degree more complex formal discourse, i.e., subjects on the level of panel discussions, news programs, etc.

LEVEL 4

SPEAKING - Subject Matter: Although the subject matter that the examinee is able to handle at this level may not differ very much from that of Level 3, he or she is able to use the language in all nontechnical situations and can express opinions almost as fully and correctly as in native language (assuming that the individual is a 5 in the native language). The examinee is able to tailor his or her speech to the audience, has near-perfect grammar and speaks the language with extensive and precise vocabulary. Although the examinee may still have an accent, he or she very rarely mispronounces the language. UNDERSTANDING: The examinee can understand the content of all conversations and formal presentations within the range of his or her experience. With the exception of dialect variations and colloquialisms outside the range of experience, understands the type of language heard in speeches sprinkled with idioms and stylistic embellishments.

LEVEL 5

SPEAKING: The examinee is able to use the language in all conceivable nontechnical subjects in a manner equivalent to that of an educated or well-informed native speaker of the language. UNDERSTANDING: The examinee is able to understand all types of formal and informal speech in a manner equivalent to that of an educated or well-informed native speaker of the language, including a wide range of dialect variations, colloquialisms, and cultural references.