

**TEST RATING STANDARDS AND
RESOURCE MATERIALS FOR RATER TRAINING:
COURT INTERPRETING ORAL PROFICIENCY EXAMINATION**

**National Center for State Courts
and
Consortium for State Court Interpreter Certification**

Revised June 2001

This document was developed partly under a grant from the State Justice Institute (SJI-95-12A-B164). Points of view and opinions stated in this report are those of the project staff and do not necessarily represent the official position or policies of the State Justice Institute nor any member of the project advisory committee.

GENERAL TEST EVALUATION AND RATING GUIDELINES

These guidelines generally apply to test rating when conducted in-person or when conducted after the fact using tape-recorded performances only. They also apply whether the test has been administered in part (screening phase, followed by a final phase) or in its entirety during a single sitting. Some specific provisions of the guidelines apply only to in-person test rating, and these should be clear from the context.

Composition of test administration and evaluation teams

Each team administering a test in person usually consists of two raters and a test administrator. The test administrator must be a native English speaker, for purposes of reading test instructions and the English portion of the consecutive. The test administrator should be a representative of the testing state or the National Center for State Courts.

At least one of the raters shall be a native speaker of the language being tested. The raters must have court-interpreting experience. After initial test administrations of new language tests, whenever possible, at least one rater shall have served previously as a rater for two or more test administrations of a Consortium test.

Selection of raters shall be based upon interpreting experience, testing experience, certifications held, and recommendations from state or federal court interpreting officials.

When there are two or more rating teams, raters may be rotated among teams, as long as there is at least one native speaker of the foreign language on each team.

Requirements for and composition of rating teams that do not administer tests in person (test rating based on tape recordings only) shall be the same as those for in-person rating.

Reference Materials

Raters should consult with each other and bring different dictionaries or other recommended reference materials to be used for rater training and final scoring, if possible.

General guidelines

Above all else, evaluators must strive for complete neutrality, fairness, and objectivity.

Conflicts of Interest

A rater should not test or rate a candidate whom the rater knows if another rater is available. Participation as a trainer in an orientation workshop does not disqualify a rater, but participation in more intensive, language-specific training shall. Candidates'

names will be provided to raters in advance of a test to check for potential conflicts. If a rater or test candidate realizes at the test that the rater and candidate know one another, the test may be given by one rater for in-person grading and the candidate's tape will also be submitted to an independent second opinion, after the process is explained to the candidate and the candidate's consent is obtained for the record.

When tests are rated using tape recordings, it is reasonable to assume candidate anonymity, unless test raters are active members of the local interpreting community. When this is the case, the test administrator should take precautions to ensure that no rater is assigned to score the tape of a person whom the rater may recognize by voice alone. Test tape labels should use an ID number (usually the candidate's social security number) to identify candidates. Names should not be included on the tapes before rating has occurred.

Confidentiality

All testing experiences shall be confidential. No information about how any candidate performed may be given to anyone other than the examiners during the examination process, the staff of the National Center for State Courts, and the testing state court administrator's office.

Testing Decorum

When raters are working in-person with candidates, raters must avoid showing reactions to a candidate. Raters should not laugh at ridiculously funny mistakes, attempt to express sympathy for ignorance, raise their eyebrows, etc. Even seemingly positive facial expressions and words may be misinterpreted.

Testing materials should be held so that the candidate cannot see them. Be as unobtrusive as possible while scoring.

RATER TRAINING

Full program

Every new test rater shall participate in a full day of test rater training that includes the elements described below and is conducted by a qualified Test Rating Supervisor.

1. ***Review procedures.*** Review general test construction theory related to scoring units and how they are used, mechanics of scoring, procedures that should be followed for efficient scoring, and guidelines for completing results report form. Because of the importance of understanding the theory of scoring units and how they are used, it is desirable that one member of the rater training faculty be someone who has participated in test construction. *It is critical that the person conducting the training does understand the theory of scoring units, how they relate to the scoring unit dictionary, and how the dictionary is to be applied, and maintained as part of the test rating process.*

2. **Test content review.** Various members of the groups read the entire test out loud.
3. **Scoring Unit and dictionary review.** Discussion of every scoring unit and the concept behind it. Review acceptables and unacceptables. The Scoring Unit Suggestion Forms should be introduced and procedures for using them explained. It is possible that rater training will surface new dictionary observations for acceptables and unacceptables, and these should be recorded on Scoring Unit Suggestion Forms. Raters should be advised that before suggested dictionary changes are actually incorporated in the official dictionary the proposed changes must be reviewed by other test raters for validation. The dictionary entries should reflect the consensus opinions of expert raters from different “groups” of raters or individuals from different parts of the country. Raters should understand that the dictionary reflects the judgment of many experts like themselves.

During training the following points should be emphasized.

- The scoring dictionary supplied with each test is a dynamic document that usually changes some after every test rating session.
- Sometimes there is no time between one administration and the next to evaluate and respond to changes recommended by earlier teams of raters. *Remember: unless the change is patently obvious (e.g., a typo), suggestions by one group of raters need to be evaluated by a larger group for consensus opinion.*
- Some suggested changes are not incorporated when the experts disagree among themselves.

Good practice for test raters: respect your colleagues. When you discover that you have a difference of opinion with another expert who contributed to the scoring dictionary, first ask yourself what reasons your colleagues may have had for what they did. We have found that rating team opinion varies one context to another, and that discussion and a wider view often makes a difference in consensus opinion.

Good rule of thumb to remember: when test raters consistently experience disagreements about how to score responses to a specific unit, it probably means that the unit is not a good one. Discovery of hidden problems with units does not occur except through test use. These units should be changed as soon as practicable and it is very important that this information be communicated in the “suggestion forms.” Discussion example: In one test, in the phrase “He hit her with a lead pipe”, test writers selected the word “lead” as a good general vocabulary scoring unit. After many problems and disagreements during scoring, it became obvious that the problem was that “lead pipe” in English isn’t used to denote something made of lead. It applies to any (heavy) metal pipe; and, in fact, what we call “lead pipes” are actually steel pipes. Thus, in that linguistic context, a very good interpreter might not use the term “lead” in the target language at all.

4. **Nonkeyword assessment review.** Discussion of nonkeyword assessment policy and the concept behind it. Practice scoring as part of review.

5. **Practice tests from tapes.** Raters shall score tapes from previous candidates as a group and fill out scoring sheets. Tapes used for training should have scores that are in the “borderline” range, where variations in rater interpretations can make the difference between a “pass” and a “fail” result for the candidate. It is important to stress that test fairness depends on the consistency (reliability) of rater opinion among different teams of test raters. The group leader will conduct “group scoring” and discussion of the entire test. Four supervised practice sessions should be included before raters begin to work in two-person teams that include new eligible raters. At least two previously scored “borderline” tapes should be scored. Supervised group rating of previously unscored tests may be included in the practice rating sessions.

Refresher training

In addition to the initial rater training, raters should participate in “refresher” training that lasts for approximately ½ day for test version being used. Refresher training may be limited to changes in the scoring dictionary if the raters have scored the test version several times previously.

SUPERVISION

One member of the test rating group should be a Test Rating Supervisor.

SCORING

CRITERIA FOR ACCEPTING RENDITIONS OF SCORING UNITS¹

A: Grammar/Verbs

If the scoring unit is a verb form, then it should be the same verb form in the target language where such exists and is appropriate usage. There is no latitude here. Example: “...parece haber ocurrido...” must be rendered as “it seems to have happened” (or “occurred”), not “it happened” or something similar. The same is true for grammatical constructions: e.g., if the scoring unit is one of gender and number, then each of the two elements must be precise in the target language. If we are testing for a rendering of a subjunctive form, that’s what it must be: e.g., if the sentence is “Lo único que quiero es que me diga lo que pasó esa noche”, it would **not** be correct to say: “All that I want to know is what happened that night.” (The subjunctive has been totally omitted; the testimony has thus been altered, even though the basic idea is there.)

Another example: “When I was a child, they would take me to the beach” would be rendered as “Cuando de niño/cuando era niño me llevaban a la playa,” not “cuando era niño me llevarían....”

¹ Although these examples and discussion have been developed for the most frequently used (Spanish) test, similar applications of these criteria are to be used for other languages, where appropriate.

B: False Cognates, etc.

There is some latitude here if the candidate does not fall for the false cognate, yet gives an acceptable rendering in the target language, even though it is less than ideal. For example, if in a Spanish test the unit is "right mind" (as in "Anyone in his right mind..."), obviously "mente correcta" is exactly what should **not** be said. However, "mente normal" could work, as could "bien de la mente", even though "en su sano juicio" or "en sus cabales" would be ideal.

C: General Vocabulary

Must be pretty much on the mark. An exception would be, for example, if the phrase is "Try to understand what was going on..." in a closing argument to a jury, and the examinee says, "Trate de" or "Intente" (i.e., singular instead of plural), it would be accepted, as the lexical item is there. Another example would be if "arma" is a scoring unit, it must be rendered as "weapon" not "gun." Sometimes usage dictates whether or not a lexical item is acceptable. For example, "illegal immigrants" would perhaps best be rendered as "inmigrantes indocumentados." However, common usage has them as "ilegales" or even "inmigrantes ilegales" which would thus both have to be accepted.

D: Legal Language, etc.

Must be precise. No room here. "Beyond a reasonable doubt" must be just that in the target language (some possibilities: "más allá de una duda razonable," "fuera de una duda razonable," "sin que quede una duda razonable"). "Testimony" must be precisely rendered in the target language, etc.

E: Idiomatic Expressions

Considered correct if: a) an equivalent idiomatic expression is given in the target language, or b) the source language idiomatic expression is recognized by the examinee and is expressed correctly but not as an idiomatic expression. This is an area in which prosodic renditions could suffice, e.g., "¡Ya lo creo" could be rendered as an emphatic "Yes!"

F: Register

No room here when there is a same-level register word in the target language. If the scoring unit is "bitch", then the rendition must be "puta", not "mala mujer" or any word that would sanitize. Another example: if the scoring unit is "uh-huh" the rendition must be "Ajá", equivalent of "yeah." If someone says "Ajá," "yes" is incorrect.

G: Numbers and Names

Must be precise. If the address is 1234 Smith Street and the examinee says, "1234 Smith Avenue", it is wrong. Dates must be exact. When referring to a year, the last two numbers are acceptable, e.g., "99" is interchangeable with "1999"; "01" with "2001."

H: Markers/Intensifiers

Here it is important that the facts conveyed by the qualifying word or the intensifier are preserved. For example: if the speaker says, "he was walking very quickly," and the scoring unit is very, the candidate must render "very" (in Spanish 'muy'), and not say "quickly" alone. However, if the speaker says, "...the earlier incident on the highway" in which "earlier" is the scoring unit, and the examinee says, "...el incidente que pasó en la carretera..." the "earlier" is implicit in the past tense and would be accepted (the incident that happened on the highway).

I: Position

Lots of latitude here. If the item is a word or phrase that might be left out because of its position, as long as something "ballpark" is there, it would be accepted. Example: So then what happened? We've accepted (for the scoring unit "So"): "Y", "Pero", "Así que", etc. (the equivalent of "and" or "but" in languages other than Spanish).

J: Slang

The same rationale as for Idiomatic Expressions would apply here. For example: If the English is "to croak," croak must be rendered with an equivalent colloquial expression or convey the meaning "to die," but not colloquially).

As a footnote, sometimes renditions that are very close but a shade off the mark may be considered "slips of the tongue" in the simultaneous portion of the test. An example would be if the examinee says "balísticas" instead of "balística", it might be considered a slip of the tongue and one could give the examinee credit for the scoring unit.

NONKEYWORD ASSESSMENT

A. Introduction

In addition to evaluation of performance by determining the number of correct scoring units, candidates are evaluated on a nonkeyword component. This is a structured assessment of evidence of language and Interpreting Skills that may not be captured within the framework of scoring units scoring. The nonkeyword assessment encompasses three dimensions: English Language Skills, Foreign Language Skills, and Interpreting Skills. Evaluation of nonkeyword performance requires the test raters to assign one of three values to the candidate's performance on each dimension. The values are **Acceptable**, **Borderline**, or **Unacceptable**.

B. Rationale for the nonkeyword assessment

Inclusion of the nonkeyword component of the overall candidate performance assessment is called for because of two types of phenomena. Experience has shown that

these can undermine the validity of scoring unit scoring as the sole evaluative mechanism.

The first circumstance is the rare case where luck has permitted a candidate to "hit" the correct interpretation of scoring units enough times to achieve the minimum acceptable score (70% or better), while routinely misinterpreting the entire unit of meaning within which the scoring unit has occurred.

The second circumstance arises when a candidate has an acceptable technical knowledge of the language (e.g., grammar, syntax, usage, vocabulary, etc.) but has severe problems speaking it -- problems that are so severe that an ordinary speaker of the language could not understand the candidate without constant interruptions and repetitions. Test raters, because of their extreme concentration, excellent listening skills, and very often shared linguistic backgrounds, may be able to understand the candidate's speech well enough to discern the intended meaning and score the test, but an ordinary listener in court would not be able to understand what is being said. A candidate may achieve a passing score on scoring units in this situation, due to the special "ear" and concentration of the raters, when it would be patently inappropriate to certify the person to court as a qualified interpreter.

C. **Impact of the nonkeyword assessment on scoring**

The nonkeyword assessment of performance functions as a corrective measure of the quantitative performance criteria associated with point scores earned through interpretations of scoring units.

Evaluators will assign an **Unacceptable** rating to performances that clearly do not meet minimum standards for court interpreting. Obviously, most **Unacceptable** ratings will be matched with scoring unit scores that do not meet the minimum standards for passing the test. However, the use of an **Unacceptable rating** on a dimension of the nonkeyword scoring system triggers consideration of failing a candidate, even if the point score is in the passing range. The procedure followed in such cases is that if both raters agree on an **Unacceptable** rating for any of the three categories, and the candidate's overall scoring unit score would otherwise entitle the candidate to pass the test, then the candidate's examination will automatically be referred to a second rating team.

NOTE: Any team rating of "unacceptable" on an examination where the candidate's point score meets the passing criterion, must be accompanied by a written statement from the raters that describes the reasons for overriding the point score. See additional explanation below.

If the second rating team also agrees on an **Unacceptable** rating on any dimension of the nonkeyword evaluation, then the candidate will not qualify for a "pass" status on the exam: regardless of his or her score on the scoring unit scoring component of the exam, the results report will be returned with a "does not pass" classification.

Assignment of an **Acceptable** score occurs when the raters believe that the interpreter's overall performance is competent or better. In such circumstances the scoring unit scoring will determine whether the candidate achieves the "pass" or "does not pass" status on the exam.

NOTE: When the candidate's point score does not meet standards for passing (70%), it is confusing and counterproductive to mark the nonkeyword evaluation as "acceptable." In effect, it says to the candidate: "The experts have considered my performance acceptable, but they are failing me anyway." Please avoid this practice as it only creates difficulties for the program managers.

A **Borderline** classification is an indication to the candidate that his/her performance on the exam demonstrated weaknesses that concerned the raters. This rating will not influence the objective (point score), so a candidate will not fail if he/she receives a borderline rating yet passes on the point score.

D. Guidelines for Scoring the Nonkeyword Component

Nonkeyword evaluations will result in assignment of one of the following rating categories for each of the three areas.

1. Scoring English Language Skills and Foreign Language Skills

The language skills criteria consider grammar, pronunciation, syntax and fluency in the language.

Acceptable rating

An **Acceptable** rating is assigned when the candidate performs very well or, at least, not alarmingly inferior. Presumptively, most candidates who pass will earn this category. Occasional errors in grammar, syntax, awkward constructions or occasional deficiencies in vocabulary or usage may be present. Accentedness that does not interfere with comprehension of the listener or the meaning of the expression may be noticed. It is appropriate to note excellent performances on the evaluation form.

Unacceptable rating

An **Unacceptable** rating in the area of language skills is assigned when a candidate's performance substantially interferes with listener comprehension or alters the meaning. Examples of unacceptable performance include recurrent grammatical errors which interfere with precise meaning, recurrent syntactical constructions that obscure understanding, and an accent that seriously interferes with a monolingual listener's comprehension or obscures the meaning of the expression.

Borderline rating

An example of **Borderline** language skills: the candidate makes several errors in syntax and grammar; the raters believe that errors are correctable with some work.

2. Scoring Interpreting Skills

The Interpreting Skills criteria consider the degree to which the candidate makes up or leaves out material, or tends to summarize; the degree to which the candidate

changes material, including gratuitous embellishments; degree to which the candidate offers alternative interpretations to what the interpretation could be and does not just get on with it; the use of appropriate pronoun forms for court settings and appropriate “voice” for references to the interpreter herself or himself.

Acceptable rating

An **Acceptable** rating is assigned in Interpreting Skills when the candidate’s Interpreting Skills are at least adequate but not inferior in any way. Examples of an adequate performance include rare occurrences of omissions, or inappropriate use of familiar voice. Room for improvement of interpreting skills may be noticed in the form of delivery that is marked by pauses and compensatory accelerations in delivery to catch up; delivery that is sometimes but not consistently choppy.

Unacceptable rating

An **Unacceptable** rating is earned when the candidate’s Interpreting Skills are clearly lacking and inappropriate. Examples of distinctively unacceptable performance include consistent changes in meaning; repeated summarizing, embellishment, omissions or paraphrasing or repeated use of inappropriate voice or person. Recurrences of “making up” material (as opposed to errors attributable to straightforward misinterpretation of words or phrases), in particular, are compelling evidence of an unacceptable level of professionalism.

Borderline rating

An example of **Borderline** Interpreting Skills: the candidate omits words or makes minor changes during the consecutive part of the test. The raters believe that the candidate has the basic skills but needs more practice to be more accurate and complete. In the simultaneous part of the test, the candidate might have shown that he/she had difficulty keeping up with the speaker.

E. Procedure for Rating the Nonkeyword Component

The last step in the nonkeyword assessment is to compare the scoring unit results scores and the nonkeyword rating to determine whether an override action is required.

If both raters have agreed on an unacceptable score for any of the three categories, and the candidate’s overall scoring unit score would otherwise entitle the candidate to pass the test, then the test raters *must complete and sign a written justification for the unacceptable rating*. This is required because the rating will override the point score result. When an override by the rating team has occurred, the candidate’s test will automatically be referred to a second rating team.

NOTE: The statement must include specific examples that illustrate the problems to the state program coordinator. The statement must be prepared as a separate document. A few sentences included in the “comments” section of the Results Report Form is not adequate documentation.

If the second team of raters also agrees on an unacceptable rating for any dimension of the nonkeyword evaluation, the candidate will fail the entire certification exam regardless of his or her score on the scoring unit component of the exam. The second team of raters must observe the same practice and procedure of rating, including completion of the written justification.

SCORING MECHANICS

Sequence of events

1. While listening to the candidate's performance live or on tape, mark the scoring units as per "marking system", below.
2. After listening to the candidate's performance on each portion of the exam and marking the scoring units as correct or incorrect, the raters should apply the nonkeyword criteria to the performance on that test part and express their judgment in the form of "A", "B", or "U" at the bottom of the last page of the script.
3. When individual scoring unit marking and the nonkeyword assessment are complete for a given test part, conduct the test rating discussion for that part, beginning with the comparative discussion of scoring units. Complete the comparison of scoring unit markings, as follows:
 - a. Designate one rater to maintain the "official" copy of the test script (lead test rater).
 - b. The lead test rater calls out the number of each scoring unit that she or he has marked incorrect, continuing to call numbers until the other rater notes a discrepancy. Each discrepancy is then discussed and resolved, and recorded on the official test script.
 - c. Scores for each page are counted and recorded at the bottom of the script - the number correct should be marked.
 - d. The total number of correct scoring units is transferred to the top of the first page of the script for that test part.
4. After recording the point scores, the raters should review their individual assessments of the nonkeyword portions of the exam and arrive at a consensus rating of "A", "B", or "U" for the performance on the test part being rated. Differences in individual evaluations should be resolved and a consensus opinion recorded on the results report form.
5. When all individual parts of the test have been rated and the results recorded on the Test Results Report Form, the section headed "Overall Comments and Summary" should be completed.
6. The Overall Evaluation (Pass/Does Not Pass) should be marked by the raters and must be consistent with the test scoring rules.

Conditions for “pass”:

- point scores must equal 70% on all three test parts
- no “unacceptable” ratings are included on any dimension of the nonkeyword scoring (note: to count as “unacceptable” on the sight interpreting subpart, both subparts must have been found to be unacceptable.)

Conditions for ‘does not pass’:

- a point score of less than 70% appears on at least one of the three test parts; or
- an “unacceptable” evaluation has been given for one or dimensions on one or more dimensions

NOTE:

So that the basis for your evaluation is as fresh as possible in your mind, and to maintain consistency among rating teams, you must complete work on each test part, including filling out the appropriate section of the results report form, before going on to listen to and score the next test part.

6. **COMMENTS:** Record observations about specific areas for improvement using the checkbox form provided. Think carefully before including other written comments. If you include your own written advisory comments, make them positive in nature. Any comments you write must **NOT** state or imply anything about the candidate’s ability to interpret in general. For example, "In this test the candidate had difficulty with vocabulary" is a defensible comment, while “The candidate lacks an adequate vocabulary” is **NOT** defensible. When it is not possible to comment favorably on an individual’s performance or to restrict comments to this performance, do not write anything.

Marking system. Mark scripts as follows:

- Incorrect interpretation of scoring unit: Put an “X” through the scoring unit number.
- Correct interpretation of scoring unit: No marks.
- Write incorrect or doubtful interpretations near the item for later reference.

We encourage (and in some controlled situations) require that each grader have two different colored pens for initial scoring and subsequent revisions. Color number one is used for the individual scoring by each rater as the test is in progress. Color number two is used to make changes.

Important: When a scoring unit originally marked "incorrect" with an "X" is changed to "correct", annotate the change by writing "OK" next to the scoring unit.

Scoring standards. A scoring unit shall be considered incorrect if:

- a. it is omitted completely or partially;
- b. the meaning is not preserved within the limits of English or the target language;
- c. grammatical precision, especially with respect to verbs, is not preserved within the limits of English or the target language; or
- d. the word or phrase has been designated as unacceptable in the scoring word list.

If a term is unfamiliar to both raters and is not in the scoring unit dictionary, check published dictionaries. If it appears in a published dictionary under the intended meaning, credit should be given. If no decision can be made, the benefit of the doubt shall be given to the candidate.

Questions regarding scoring units.

When questions or concerns regarding any scoring unit arise during the administration of the exam, raters should make a note of the units in question and discuss at a group meeting at the end of the day. If there is only one team, the question should be submitted to the National Center for State Courts on the forms provided.

These suggestions regarding scoring units and scripts should be recorded on the “Scoring unit Suggestion report” form and returned to the NCSC with all the other material. **This is an important part of raters’ work and enough time should be provided for in contracts to complete the process.** It is the responsibility of the lead test rater or the on-site rating contract supervisor to see that this is done.

Group meeting.

If more than one team is involved, a meeting at the end of the day will help to discuss issues regarding scoring and adjustments to scoring units.