

TEST CONSTRUCTION MANUAL

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

**REVISED EDITION
NOVEMBER 2000**

This document was developed 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.

Test Construction Manual

1.0 MANAGEMENT OF THE TEST DEVELOPMENT

Under the general oversight of the Technical Committee and Consortium staff, the development of each test should be overseen and managed by a Test Development Manager. The Test Development Manager should have experience writing court interpreter examinations. He or she should also understand and follow the Consortium's policies for test development articulated in the Manual.

The Test Development Manager is responsible for assembling and supervising the test development team as well as the legal, linguistic, and cultural reviewers and ensuring that the team and reviewers perform their duties in a manner consistent with this Manual. The Test Development Manager should solicit from Consortium members and other experts suggestions for members of test development teams as well as reviewers, transcripts and other documents that might serve as base texts for tests, and other forms of assistance, advice, and resources as the Consortium deems appropriate.

2.0 SELECTION OF BASE DOCUMENTS

1. The base documents which serve as the starting point for developing the consecutive or simultaneous component of an examination should be selected from transcripts of official court proceedings from:
 - A. A wide range of stages of legal proceedings (e.g., preliminary hearings, trials, post-trial activity such as sentencing, etc.).
 - B. Proceedings of all types of legal subjects heard in state courts (e.g., criminal, family, and civil courts).
 - C. Proceedings of all levels of non-Federal courts (e.g., state, county, municipal).
2. The following types of proceedings should be avoided, except when the transcripts can easily be edited to meet Consortium standards:
 - A. Proceedings whose discourse is highly technical (e.g., expert witness testimony, sophisticated legal argumentation).
 - B. Proceedings with discourse that contains an inordinate number of legal terms (e.g., motions and legal arguments on technical points of law). The number of legal terms that are not scoring units should not exceed the number of legal terms that are scoring units in the final base text of any test.
3. The base documents which serve as the starting point for developing the sight component of an examination should be selected from documents actually used by a court such as:
 - A. Police and other law enforcement reports. (for English to foreign language sight)
 - B. Other reports (e.g., mental health) likely to be included with an investigative report (e.g., pre-sentence report, child custody report) prepared for a court. (for English to foreign language sight)
 - C. Correspondence to judges and other court staff employees (e.g., character reference letters to judges). (for foreign language to English)
 - D. Any investigative report prepared for a court to which parties are entitled access (e.g., pre-sentence or pre-disposition report). (for foreign language to English)

Test Construction Manual

- E. Publications such as journals or publications not widely known or easily available.
- 4. Documents whose text is more or less established and non-variable should be avoided since they are readily available in the public domain and that availability can compromise test security. Examples include:
 - A. Forms.
 - B. Form letters.
 - C. Standard statements that do not vary in text (e.g., standard conditions of probation).
- 5. All materials selected should be representative types of material and not isolated, specialized, or arcane.
- 6. All materials selected should be from types of legal situations where there are large volumes of interpreting services required.
- 7. All materials selected should involve general legal concepts and procedures and avoid unique state proceedings, except where they can be easily adapted to meet Consortium standards.
- 8. Over time, the Consortium will increase the use of civil and family transcripts, subject to #6, while maintaining a significant use of criminal transcripts.

3.0 SCORING UNITS

Tests are scored on the basis of scoring units which are special linguistic characteristics that interpreters must be able to render to deliver a complete and accurate interpretation. Operationally, a scoring unit is a pre-selected portion of the exam material that is underlined in a rater's transcript of the test text. The following types of units will be included. Frequencies of each type will be as specified in 5.0, Distribution of Scoring Units in Entire Test, except as specifically modified by expert language consultants for each new language added to the test battery. The reasons for the modifications, if any, will be described and recorded in a memorandum to the Consortium Technical Committee.

1. Grammar and Usage

- A. Grammar/Verbs
Features of grammar, especially verbs, that may not be handled appropriately by the unsophisticated user of the two languages.
- B. False Cognates/Interference/Literalism
Terms or phrases that are likely to invite misinterpretation due to interference by one language on the other, e.g., false cognates, awkward phrasing; terms or phrases susceptible to literal renditions resulting in loss of precise meaning.

Test Construction Manual

2. General Lexical Range

C. General Vocabulary

Any general lexical item or set of items not easily classified elsewhere among the scoring units.

D. Legal Terms & Phrases

Any legal word or phrase of a legal or technical nature, or which is not common in everyday speech but is commonly used in legal settings.

E. Idioms/Sayings

“Idioms” are sets of words whose meaning as a whole is different from the meaning of the individual words.¹ “Sayings” includes famous sayings from literature, history, etc.

3. Conservation of More Technical Forms

F. Register

Words and phrases of unquestionably high or low register that can be preserved in the target language, but might be lowered or raised (e.g., curses, profanity, taboo words).

G. Numbers/Names

Any number (e.g., street address, weight of person or object, measurements such as distance) or name (e.g., person, court, street, town).

H. Markers/Intensifiers/Emphases/Precision

Any word or phrase giving emphasis or precision to a description (e.g., adverbs, adjectives) or statement (e.g., can be grammatical in form).

I. Embeddings/Positions

Word or phrases likely to be omitted due to position (e.g., at the beginning or in the middle of a long sentence; the second in a string of adjectives or adverbs) or function (e.g., tag questions).

J. Slang/Colloquialisms

Words/phrases that seem to be slang or colloquial language.²

4.0 GENERAL PHILOSOPHY OF SELECTING AND CLASSIFYING SCORING UNITS

1. The meaning of the source language unit must be unambiguous.
2. The ways the scoring unit can be rendered in the target language should be clear and beyond debate due to dialect differences. If there is dialectal variety, then either all dialects must be permitted as acceptable or the prospective scoring unit should not be chosen as a scoring unit.
3. The various scoring units should be distributed throughout the test’s components so that their distribution across the entire test is consistent with the ranges established elsewhere herein.
4. A scoring unit may consist of any word or group of words.
5. A scoring unit may or may not be a critical part of the discourse.

¹ There are two technical definitions of “idiom”: “Sequence of words which is semantically and often syntactically restricted, so that they function as a single unit.” David Crystal, *A DICTIONARY OF LINGUISTICS AND PHONETICS*, 2nd Ed. 152 (1985). “Fixed phrases, consisting of more than one word, with meanings that cannot be inferred by knowing the meanings of the individual words.” Victoria Fromkin and Robert Rodman, *AN INTRODUCTION TO LANGUAGE*, Third Ed. 181 (1983).

² “Slang” is difficult to define. For reasons why and examples, see Fromkin and Rodman, *id.*, at 264-265.

c:\documents and settings\al2smith.olympus\local settings\temporary internet files\olk9\web 13 oe construction manual.doc

Test Construction Manual

6. The distribution of scoring units within each type should include units of varying estimated degrees of difficulty.
7. A particular form of any scoring unit should appear only once in a given examination. For example, a particular verb form should appear as a grammatical scoring unit only once in the test and should not also appear elsewhere in that test as either some other form of the same verb or general vocabulary.
8. When the choice between two classifications could go either way, decide whether it is worth keeping the proposed scoring unit at all. If the decision is to keep the scoring unit, classify it as one type and be clear that it has been selected for that testing purpose.
9. When the form of the word(s) in a proposed scoring unit is not important and when that proposed scoring unit is not already clearly some other kind of scoring unit, that proposed scoring unit is a good candidate to be considered for the General Vocabulary category.
10. Specific types of scoring units should follow as much of a systematic structure as possible. Examples of such structure are given below for two types of scoring units. The Technical Committee of the Consortium will be responsible for developing examples of the structural requirements of the other types of scoring units.

A. Grammar/Verbs

The first test-writing team in a new language should prepare a list of the major problems of grammar and verbs for each language before the test development process begins. For example, in Spanish agreement in terms of gender and number between adjectives and nouns is important. Also important in Spanish is appropriate use of subjunctive form. Another example is appropriate use of ser and estar.

Once a list of the major features has been agreed upon by experts in the respective languages, then test developers can make sure they are represented as broadly as possible.

G. Numbers/Names

As many of the following types of numbers should be included as possible:

1. Street/ mailing addresses
2. Weight
3. Distances
4. Dates
5. Time of day
6. Ages
7. Legal processing numbers (e.g., docket numbers, exhibit numbers)
8. Telephone numbers
9. Dollar amounts/currency
10. Personal identification numbers (e.g., social security, drivers' license)

As a general rule, there should be no more than one of each of these types in the same test within a language. However, the same type could appear once in English as the source language as well as once in the other source language.

11. Suggested steps for identifying scoring units:

Test Construction Manual

Step One: Pass through a text and assign scoring units based on first impressions.

Step Two: Calculate the types of scoring units that are generated on the first run and identify those which are over-represented and those which are under-represented in each test part.

Step Three: For those that are over-represented, select the ones that appear best to keep (based on perceived strength of the scoring units themselves as well as on their proximity to other scoring units) and eliminate as many as may be necessary to approximate the target number for each type.

Step Four: For those that are underrepresented, edit the text in such a fashion as to add material that includes these types of scoring units.

12. Where possible, specific existing dictionaries can be identified as a source for certain types of scoring units. An example would be dictionaries of idioms, slang, and/or colloquialism. A list should be developed by linguistic experts within the language and be made available to the test writers for that specific language.
13. Care should be taken to distribute scoring units throughout texts to avoid clustering.

5.0 DISTRIBUTION OF SCORING UNITS IN ENTIRE TEST

The scoring units should be distributed among the test components so that each test has the following distribution overall (+/- 10% within major categories acceptable):

Grammar and Usage:		25%
Grammar/Verbs:	15%	
Interference:	10%	
General Lexical Range:		40%
General Vocabulary:	20%	
Legal Terms/Phrases:	15%	
Idioms:	5%	
Conservation of More Technical Forms:		35%
Register:	5%	
Numbers:	7%	
Markers:	10%	
Position:	9%	
Slang:	4%	

So far as is possible, the various types of scoring units should appear in all components of the test. **Table 1** shows the recommended distribution of units for each test part. The test development team should review the appropriateness of these percentages for each language combination and recommend adjustments to them when the linguistic properties of English and the second language make an adjustment necessary.

Test Construction Manual

TABLE 1 – SCORING UNIT DISTRIBUTION

Whole Test	S-E	S-F	Con	Sim	Sum	Target %
A	4	4	15	10	33	15
B	3	3	9	6	21	10
SUB	7	7	24	16	54	25
C	8	8	15	13	44	20
D	3	3	11	16	33	15
E	0	0	7	4	11	5
SUB	11	11	33	33	88	41
F	1	1	5	3	10	5
G	1	2	6	5	14	7
H	3	3	9	7	22	10
I	1	1	9	8	19	9
J	1	0	4	3	8	4
SUB	7	7	33	26	73	34
TOTAL	25	25	90	75	215	100

6.0 DISTRIBUTION OF SCORING UNITS IN THE CONSECUTIVE COMPONENT

When drafting the consecutive component of the examination, the utterances to be interpreted should be of varied lengths ranging from one word to a high no greater than fifty words. No more than 20 words should be used in the first two utterances of the consecutive test and no more than 30 words in the first four utterances. Frequency and/or difficulty of scoring units should also be relatively sparse in the initial four utterances. The scoring units should be embedded in utterances that vary in length approximating the distribution shown in **Table 2**.

TABLE 2

UTTERANCE LENGTH (In Number of Words)	DISTRIBUTION OF SCORING UNITS IN THE CONSECUTIVE PER SOURCE-TARGET LANGUAGE SEGMENT	
	English → Foreign Language	Foreign Language → English
1-10	10%	10%
11-20	25%	25%
21-30	30%	30%
31-40	25%	25%
Subtotal 11-40**80% (+/- 10%)	80% (+/- 10%)
41-50	10%	10%
Total	100%	100%

** The range of 21-30 must always be greater than either the 11-20 range or the 31-40 range; the subtotal of 11-40 may have a variation of up to 10%.

Test Construction Manual

7.0 LENGTH OF TEST COMPONENTS

The size of each test component should be within the ranges shown in **TABLE 3**.

TABLE 3

COMPONENT OF TEST	RANGE PER NUMBER OF WORDS
Sight	400-450
English → Foreign Language	200-225
Foreign Language → English	200-225
Consecutive	850-950
English → Foreign Language	425-475
Foreign Language → English	425-475
Simultaneous	800-850
Totals	1950-2250

8.0 NUMBER AND WEIGHTING OF SCORING UNITS

Each test should contain 215 scoring units distributed among the various components as indicated in **TABLE 4**, below. In addition, calculations of the overall average score should be based on the weights identified in the table.

TABLE 4

COMPONENT OF TEST	TOTAL NUMBER OF SCORING UNITS	WEIGHT
Sight	50	20%
English → Foreign Language	25	
Foreign Language → English	25	
Consecutive	90	40%
English → Foreign Language	Approximately 40	
Foreign Language → English	Approximately 50	
Simultaneous	75	40%
Totals	215	

9.0 DOCUMENTATION OF TEST DEVELOPMENT PROCESS

Staff of the Consortium should collect and preserve information about the development of each test, including the names and qualifications of every person involved.

10.0 PILOT TESTING

Each version of every test should be trial tested before it is given to the general public. Trial testing is used as a simulation of the operational test administration and scoring to identify and correct problems prior to the first operational administration and scoring. Among other things, trial testing identifies additional acceptable and unacceptable responses, items that may

Test Construction Manual

not function as intended, and some items that pose problems for raters. When possible, the following guidelines should be followed when preparing a new test version for pilot testing.

1. The pilot test should be given in a minimum of two geographically separate parts of the United States to as many different speakers of the language by regional varieties (if applicable) as possible, focusing primarily on the groups of speakers who reside in the United States generally or in the member states of the Consortium.
2. If possible, the trial test should be given to approximately three persons in each of the following groups:
 - A. Persons with a high probability of scoring very well based on performance on other tests (over 75%);
 - B. Persons with a probability of scoring at or near the minimum acceptable level for passing; and
 - C. Persons who are novices or students.
3. After the trial test is administered, the results should be reviewed in general and in the following specific respects:
 - A. Scoring units that are difficult to grade since they are packed too tightly together.
 - B. Words or phrases that give examinees unanticipated problems and require editing of the general text, not necessarily scoring units themselves.

11.0 STATISTICAL ANALYSIS OF TEST

After an adequate sample of examinees has taken a given test version, Consortium staff should arrange an analysis of the test to address reliability and any basic validity issues that may arise. Any items with negative discrimination will be eliminated from future test versions. Information obtained through the analysis that points to low discrimination power of any item will be used to revise or delete that item from future test versions.

12.0 QUALIFICATIONS OF TEST DEVELOPMENT TEAMS

Each test should be written primarily by a team consisting of at least two specialists. The team should consist of one practicing professional interpreter (or interpreter/translator) with the highest credentials available in the field and the best possible complement of specialists available from the following categories:

- (1) Practicing professional interpreter (or interpreter/translator) with the highest credentials available in the field;
- (2) Academicians with the most formal training possible in the linguistics of the language; and
- (3) Bilingual professionals (e.g., attorneys, teachers) who are native speakers of the language.

When it is not possible to assemble a team as described above, the team should consist of persons who are otherwise deemed to be the most competent for the purpose of writing a test.

Test Construction Manual

13.0 LEGAL REVIEW

Prior to testing, every base text of each examination should be reviewed by at least one attorney who practices law in three different member states. Their review will focus on the following:

1. Appropriateness of any legal terms for use in their respective jurisdictions and, if known, other jurisdictions as well. The goal is to use terms that are fairly universal and to avoid terms that are unique or regional.
2. Appropriateness of the discourse of the text in terms of whether it is consistent with acceptable legal practice and procedure.

14.0 CULTURAL REVIEW

Prior to testing, every new test should be reviewed by at least one expert of the language's culture(s) for cultural appropriateness. The aims are:

1. To make sure that the test content is culturally appropriate for the group of people who will benefit from interpreting in the given language (e.g., person from the group have been known to engage in the conduct that is the subject of the proceeding); and
2. To make sure that there are no other reasons for offending persons from the cultural group in question due to insensitivity or sociolinguistic gaffes.

15.0 LINGUISTIC REVIEW

Prior to testing, every test for a given language should be reviewed by professional linguists or interpreters representing different dialects of the language. It is important to produce a text that is as dialect-free as possible, or at least mutually intelligible across all dialects.

16.0 PRODUCTION OF THE SIMULTANEOUS TAPE OR CD

The recording by which the simultaneous or consecutive components are administered should be produced as follows:

1. The recording should be produced according to professional standards of sound recording so that it is clear, free from extraneous noises (including hiss), and easily understood.
2. The tape or CD should be recorded at a constant speed of 120 words per minute, as far as is possible.
3. No single minute should be slower than 110 words per minute or faster than 130 words per minute.
4. No scoring units should appear in the first ten seconds of the recording in order to give the examinee a chance to get going.
5. The speakers should strive to emulate the quality of broadcast voice.

Test Construction Manual

17.0 MISCELLANEOUS CONSIDERATIONS

Unit of Count

“Word” should be counted as follows:

- A. Anything that would be separated by a space or punctuation when written constitutes one word.
- B. Each hyphenated portion of a word that would be hyphenated when written counts as one word (e.g., “twenty-three” counts as two words).
- C. Numbers should be counted the way they would be written out as words (e.g., “1995” would be “nineteen, ninety-five” counts as three words).
- D. The words being counted are words in the source language, not the target language.