



# **Versant<sup>™</sup> Arabic Test**

Test Description and Validation Summary

## Table of Contents

<b>Section I - Test Description</b>	<b>0</b>
<b>1. Introduction</b>	<b>0</b>
<b>2. Test Description</b>	<b>3</b>
2.1 Modern Standard Arabic	3
2.2 Test Design	4
2.3 Test Administration	4
2.3.1 Telephone Administration	5
2.3.2 Computer Administration	5
2.4 Test Format	5
Part A: Readings	5
Parts B and E: Repeats	6
Part C: Short Answer Questions	7
Part D: Sentence Builds	8
Part F: Passage Retellings	8
2.5 Number of Items	9
2.6 Test Construct	9
<b>3. Content Design and Development</b>	<b>12</b>
3.1 Rationale	12
3.2 Vocabulary Selection	12
3.3 Item Development	13
3.4 Item Prompt Recording	13
3.4.1 Voice Distribution	13
3.4.2 Recording Review	14
<b>4. Score Reporting</b>	<b>14</b>
4.1 Scores and Weights	14
4.2 Score Use	16
<b>Section II – Field Test and Validation Studies</b>	<b>16</b>
<b>5. Field Test</b>	<b>16</b>
5.1 Data Collection	16
5.1.1 Native Speakers	17
5.1.2 Non-Native Speakers	17
<b>6. Data Resources for Score Development</b>	<b>17</b>
6.1 Data Preparation	17
6.1.1 Transcription	17
6.1.2 Human Rating	18
<b>7. Validation</b>	<b>18</b>
7.1 Validity Study Design	18
7.1.1 Validation Sample	19
7.1.2 Test Materials	19

7.2 Internal Validity.....	20
7.2.1 Validation Sample Statistics.....	20
7.2.2 Test Reliability .....	20
7.2.3 Dimensionality: Correlations between Subscores.....	21
7.2.4 Machine Accuracy: VAT Scored by Machine vs. Scored by Human Raters .....	22
7.2.5 Differences among Known Populations.....	23
7.3 Concurrent Validity: Correlations between VAT and Human Scores.....	25
7.3.1 Concurrent Measures.....	25
7.3.2 OPI Reliability.....	27
7.3.3 VAT and ILR OPIs.....	28
7.3.4 VAT and ILR Level Estimates .....	28
7.3.5 VAT and CEFR Level Estimates.....	30
<b>8. Conclusion .....</b>	<b>32</b>
<b>9. About the Company.....</b>	<b>33</b>
<b>10. References .....</b>	<b>34</b>
<b>11. Textbook References .....</b>	<b>35</b>
<b>12. Appendix: Test Materials .....</b>	<b>37</b>

## Section 1 – Test Description

### 1. Introduction

Pearson's Versant™ Arabic Test (VAT), powered by Ordinate technology, is an assessment instrument designed to measure how well a person understands and speaks Modern Standard Arabic (MSA). MSA is a non-colloquial language, which is deemed suitable for use in writing and in spoken communication within public, literary, and educational settings. The VAT is intended for adults and students over the age of 18 and takes approximately 17 minutes to complete. Because the VAT test is delivered automatically by the Ordinate testing system, the test can be taken at any time, from any location by phone or via computer and a human examiner is not required. The computerized scoring allows for immediate, objective, reliable results that correspond well with traditional measures of spoken Arabic performance.

The Versant Arabic Test measures *facility* with spoken Arabic, which is a key element in Arabic oral proficiency. Facility with MSA is how well the person can understand spoken Modern Standard Arabic on everyday topics and respond appropriately at a native-like conversational pace in Modern Standard Arabic. Educational, commercial, and other institutions may use VAT scores in decisions where the measurement of listening and speaking is an important element. VAT scores provide reliable information that can be applied in placement, qualification and certification decisions, as well as in progress monitoring or in the measurement of instructional outcomes.

### 2. Test Description

#### 2.1 Modern Standard Arabic

Different forms of Arabic are spoken in the countries of North Africa and the Middle East, extending roughly over an area from Morocco and Mauritania in the west, to Syria and Iraq in the northeast, to Oman in the southeast. Each population group has a colloquial form of Arabic that is used in daily life (sometimes along with another language, e.g. Berber or Kurdish). All population groups recognize a non-colloquial language, commonly known in English as Modern Standard Arabic (MSA), which is suitable for use in writing and in spoken communication within public, literary, and educational settings.

Analyzing a written Arabic text, one can often determine the degree to which the text qualifies as MSA by examining linguistic aspects of the text such as its syntax and its lexical forms. However, in spoken Arabic, there are other salient aspects of the language that are not disambiguated in the usual form of the written language. For example, if a person reads aloud a short excerpt from a newspaper but vocalizes with incorrect case markings, one might conclude that the reader does not know case rules. Nevertheless one would not necessarily conclude that the newspaper text itself is not MSA. Also, in phonological terms, native speakers of Arabic can be heard pronouncing specific words differently, depending on the speaker's educational or regional background. For example, the MSA demonstrative *هَذَا* /haḏa/ is frequently uttered as /haza/. The speech of Arabs on radio and television includes a wide

variation in the syntax, phonology, and lexicon within what is intended to be MSA. Thus, the boundaries of MSA may be clearer in its written form than in its several spoken forms.

## 2.2 Test Design

The VAT may be taken at any time from any location using a telephone or a computer. During test administration, the Ordinate testing system presents a series of recorded spoken prompts in Arabic at a conversational pace and elicits oral responses in Arabic. The voices that present the item prompts belong to native speakers of Arabic from several different countries, providing a range of native accents and speaking styles.

The VAT has five task types that are arranged in six sections: Readings, Repeats (presented in two sections), Short Answer Questions, Sentence Builds, and Passage Retellings. All items in the first five sections elicit responses from the test-taker that are analyzed automatically by Ordinate scoring system. These item types provide multiple, fully independent measures that underlie facility with spoken MSA, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, listening skill, and pronunciation of rhythmic and segmental units. Because more than one task type contributes to each subscore, the use of multiple item types strengthens score reliability.

The VAT score report is comprised of an Overall score and four diagnostic subscores:

- Sentence Mastery
- Vocabulary
- Fluency
- Pronunciation

Together, these scores describe the test-taker's facility in spoken Arabic. The Overall score is a weighted average of the four subscores.

The Ordinate testing system automatically analyzes the test-taker's responses and posts scores on its website within minutes of completing the test. Test administrators and score users can view and print out test results from a password-protected section of Pearson's website.

## 2.3 Test Administration

Administration of a VAT test generally takes about 17 minutes over the phone or via a computer. Regardless of the mode of test administration, it is best practice (even for computer delivered tests) for the administrator to give a test paper to the test-taker at least five minutes before starting the VAT test. The test-taker then has the opportunity to read both sides of the test paper and ask questions before the test begins. The administrator should answer any procedural or content questions that the test-taker may have.

The mechanism for the delivery of the recorded item prompts is interactive – the system detects when the test-taker has finished responding to one item and then presents the next item.

### 2.3.1 Telephone Administration

Telephone administration is supported by a test paper. The test paper is a single sheet of paper with material printed on both sides. The first side contains general instructions and an explanation of the test procedures (see Appendix). These instructions are the same for all test-takers. The second side has the individual test form, which contains the phone number to call, the Test Identification Number, the spoken instructions written out verbatim, item examples, and the printed sentences for Part A: Reading. The individual test form is unique for each test-taker.

When the test-taker calls the Ordinate testing system, the system will ask the test-taker to use the telephone keypad to enter the Test Identification Number that is printed on the test paper. This identification number is unique for each test-taker and keeps the test-taker's information secure.

A single examiner voice presents all the spoken instructions for the test. The spoken instructions for each section are also printed verbatim on the test paper to help ensure that test-takers understand the directions. These instructions (spoken and printed) are available either in English or in Arabic. Test-takers interact with the test system in Arabic, going through all six parts of the test until they complete the test and hang up the telephone.

### 2.3.2 Computer Administration

For computer administration, the computer must have an Internet connection and Ordinate's Computer Delivered Test (CDT) software. It is best practice to provide the test-taker with a printed test paper to review before the actual computer-based testing begins. The test-taker is fitted with a microphone headset. The CDT software requires the test-taker to adjust the volume and calibrate the microphone before the test begins.

The instructions for each section are spoken by an examiner voice and are also displayed on the computer screen. Test-takers interact with the test system in Arabic, speaking their responses into the microphone. When a test is finished, the test-taker clicks a button labeled "END TEST".

## 2.4 Test Format

During the test administration, the instructions for the test are presented orally in the unique examiner voice and they are also printed verbatim on the test paper or on the computer screen. Test items themselves are presented in various native-speaker voices that are distinct from the examiner voice.

The following subsections provide brief descriptions of the task types and the abilities that can be assessed by analysis of the responses to the items in each part of the VAT test.

### Part A: Readings

In the Reading task, test-takers read printed, numbered sentences, one at a time, in the order requested by the examiner voice. The reading texts are printed on a test paper which should be given to the test-taker before the start of the test. On the test paper or on the computer screen, reading

items are voweled and are grouped into sets of four sequentially coherent sentences as in the example below.

Examples:

1. لَا يُحِبُّ مُحَمَّدٌ شَقَّتَهُ.

2. الازدحام شديد في الشارع أمام البيت والمياه دائماً مقطوعة عنه.

3. لذلك فهو يحاول العثور على سكن جديد.

4. ولكن كل الشقق الجديدة التي وجدها غالية جداً.

1. Mohamed does not like his apartment.
2. It's very crowded in the street in front of the house, and there's no water at all.
3. That's why he's trying to find another place to live.
4. But all the new apartments he's found are very expensive.

Presenting the sentences in a group helps the test-taker disambiguate words in context and helps suggest how each individual sentence should be read aloud. The test paper (or computer screen) presents two sets of four sentences and the examiner voice instructs the test-taker which of the numbered sentences to read aloud, one-by-one in a random order (e.g., *Please read Sentence 4. ... Now read Sentence 1. ... etc.*). After the system detects silence indicating the end of one response, it prompts the test-taker to read another sentence from the list.

The sentences are relatively simple in structure and vocabulary, so they can be read easily and fluently by people educated in MSA. For test-takers with little facility in spoken Arabic but with some reading skills, this task provides samples of their pronunciation and oral reading fluency. The readings start the test because, for some test-takers, reading aloud presents a familiar task and is a comfortable introduction to the interactive mode of the test as a whole.

## Parts B and E: Repeats

In the Repeat task, test-takers are asked to repeat sentences verbatim. Sentences range in length from three words to twelve words, although few item sentences are longer than nine words. The audio item prompts are spoken aloud by native speakers of Arabic and are presented to the test-taker in an approximate order of increasing difficulty, as estimated by item-response analysis.

Examples:

عبد الله هو الذي قال ذلك.  
الكراسي كانت أساس ربح الشركة.  
أكثر من مائة تلميذ اضطروا للبقاء في بيوتهم.

Abdulla is the one who said so.  
Chairs were the mainstay of the company's profit.  
More than one hundred students had to stay at home.

To repeat a sentence longer than about seven syllables, the test-taker has to recognize the words as produced in a continuous stream of speech (Miller & Isard, 1963). However, highly proficient speakers of Arabic can generally repeat sentences that contain many more than seven syllables because these speakers are very familiar with Arabic words, collocations, phrase structures, and other common linguistic forms. In English, if a person habitually processes four-word phrases as a unit (e.g., “the furry black cat”), then that person can usually repeat verbatim English utterances of 12 or even 16 words in length. A typical Arabic typographic word carries more morpho-semantic units, so Arabic words typically carry more information than a usual English word. For example, on average, it takes about 140 English words to translate a 100-word Arabic passage. Therefore, a 10- or 11-word limit on Arabic Repeat items should roughly correspond to the 14- or 15-word limit on the Versant English Repeat items. Generally, the ability to repeat material is constrained by the size of the linguistic unit that a person can process in an automatic or nearly automatic fashion. As the sentences increase in length and complexity, the task becomes increasingly difficult for speakers who are not familiar with Arabic phrase and sentence structure.

Because the Repeat items require test-takers to organize speech into linguistic units, Repeat items assess the test-taker's mastery of phrase and sentence structure. Given that the task requires the test-taker to repeat back full sentences (as opposed to just words and phrases), it also offers a sample of the test-taker's fluency and pronunciation in continuous spoken Arabic.

### Part C: Short Answer Questions

In this task, test-takers listen to spoken questions in Arabic and answer each question with a single word or short phrase. The questions generally include at least three or four content words embedded in some particular Arabic interrogative structure. Each question asks for basic information, or requires simple inferences based on time, sequence, number, lexical content, or logic. The questions are designed not to presume any specialist knowledge of specific facts of Arabic culture, geography, religion, history, or other subject matter. They are intended to be within the realm of familiarity of both a typical 12-year-old native speaker of Arabic and an adult learner who has never lived in an Arabic-speaking country.

Examples:



كم عيناً للإنسان ؟

إن كنت مريضاً فهل تذهب للطبيب أم للمحامي ؟

How many eyes does a human (usually) have?

If you are unwell do you go to a doctor or a lawyer?

To respond to the questions, the test-taker needs to identify the words in phonological and syntactic context, and then infer the demand proposition. Short Answer Questions manifest a test of receptive and productive vocabulary within the context of spoken questions.

### Part D: Sentence Builds

For the Sentence Build task, test-takers are presented with three short phrases. The phrases are presented in a random order (excluding the original, most sensible, phrase order), and the test-taker is asked to rearrange them into a sentence, that is, to speak a reasonable sentence that comprises exactly the three given phrases.

Examples:

مع عائلته / أن يخرج / لا يحب

تحب ابنتها / أصدقائها / تصوير

with his family / to go out / he doesn't like

her daughter likes / her friends / taking pictures of

In the Sentence Build task, the test-taker has to understand the possible meanings of each phrase and know how the phrases might combine with the other phrasal material, both with regard to syntax and semantics. The length and complexity of the sentence that can be built is constrained by the size of the linguistic unit (e.g., one word *versus* a two- or three-word phrase) that a person can hold in verbal working memory. This is important to measure because it reflects the candidate's ability to access and retrieve lexical items and to build phrases and clause structures automatically. The more automatic these processes are, the more the test-taker demonstrates facility in spoken Arabic. This skill is demonstrably distinct from memory span (see Section 2.6, Test Construct, below).

The Sentence Build task involves constructing and saying entire sentences. As such, it is a measure of the test-taker's mastery of language structure as well as pronunciation and fluency.

### Part F: Passage Retellings

In the final VAT task, test-takers listen to a spoken passage (usually a story) and then are asked to describe what happened in their own words. Test-takers are encouraged to re-tell as much of the passage as they can, including the situation, characters, actions and ending. The passages are from 19 to 50 words in length. Most passages are simple stories with a situation involving a character (or

characters), a setting, and a goal. The body of the story typically describes an action performed by the agent of the story followed by a possible reaction or implicit sequence of events. The ending sometimes introduces a new situation, actor, patient, thought, decision, or emotion.

Examples:

كان ثلاثة إخوة يسيرون في الطريق وفجأة وجدوا سيارة مقلوبة. لم يعرفوا كيف يساعدون السائق، فقرروا أن يرجعوا إلى البيت ليطلبوا المساعدة من والدهم.

Three brothers were walking down the street when suddenly they came upon an overturned car. They didn't know how to help the driver, and so they decided to return home to ask for their father's help.

Passage Retellings capture the test-taker's listening comprehension ability and also provide additional samples of spontaneous speech. Currently, this task is not automatically scored by the computerized scoring system, but response recordings are available online on the password-protected section of the website and can be reviewed by human listeners for validation purposes.

## 2.5 Number of Items

During each VAT administration, 69 items in total are presented to the test-taker in the six separate sections, Parts A through F. In each task section, the items are drawn from a much larger item pool. For example, each test-taker is presented with ten Sentence Build items selected quasi-randomly from the pool, so most or all items will be different from one test administration to the next. The Ordinate testing system selects items from the item pool taking into consideration, among other things, the item's level of difficulty and its form and content in relation to other selected items. Table 1 shows the number of items presented in each section.

Table 1. Number of items presented per task.

Task	Presented
A. Readings	6 (of 8 printed)
B. Repeats	15
C. Short Answer Questions	20
D. Sentence Builds	10
E. Repeats	15
F. Passage Retellings	3
<b>Total</b>	<b>69</b>

## 2.6 Test Construct

For any language test, it is essential to define the test construct as explicitly as possible. As presented above in Section 2.1, one observes minor variations in the spoken forms of Modern Standard Arabic. The Versant Arabic Test (VAT) is designed to measure a test-taker's *facility in spoken forms of Modern Standard Arabic* – that is, *the ability to understand and speak contemporary Modern Standard Arabic as it is used in international communication for broadcast, for commerce, and for professional collaboration*. Because a person knowledgeable in MSA needs to understand Arabic as it is currently used, e.g., on Al Jazeera broadcasts, the VAT item recordings include a range of local phonological features that reflect commonly encountered forms of the language, while adhering to prescribed vocabulary, syntax, and case forms of MSA. Thus, while some specific low-prestige segmental substitutions have been excluded from the item recordings (e.g., \*/s/ for /th/, and \*/z/ for /ḍ/), other segmental substitutions have been allowed (e.g., voicing /ṣ/ in \*/'azbaḥa/ for /'aṣḥaḥa/). For more detail, see Section 3.4, *Item Prompt Recording*.

There are many processing elements required to participate in a spoken conversation: a person has to track what is being said, extract meaning as speech continues, and then formulate and produce a relevant and intelligible response. These component processes of listening and speaking are schematized in Figure 1, adapted from Levelt (1989).

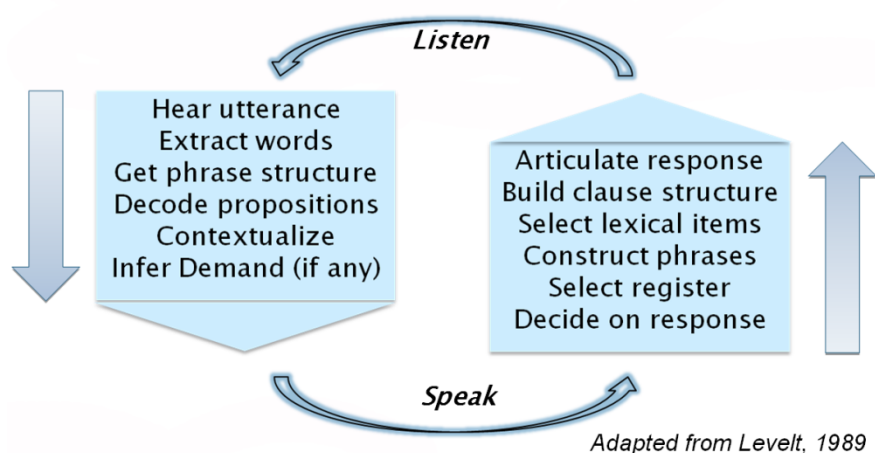


Figure 1. Conversational processing components in listening and speaking.

Core language component processes, such as lexical access and syntactic encoding, typically take place at a very rapid pace. During spoken conversation, Van Turenhout, Hagoort, and Brown (1998) found that Dutch native speakers go from building a clause structure to phonetic encoding in about 40 milliseconds. Similarly, the other stages shown in Figure 1 have to be performed within the small period of time available to a speaker involved in interactive spoken communication. A typical window in turn taking is about 500-1000 milliseconds (Bull and Aylett, 1998). Although most research has been conducted with English and European languages, it is expected that the results will closely resemble the processing of other languages including Arabic. If language users cannot perform the internal activities presented in Figure 1 in real time, they will not be able to participate as effective listener/speakers. Thus, spoken language facility is essential in successful oral communication.

Because test-takers respond to the VAT items in real time, the system can estimate the test-taker's level of automaticity with the language with respect to the latency and pace of the spoken response. Automaticity is the ability to access and retrieve lexical items, to build phrases and clause structures, and to articulate responses without conscious attention to the linguistic code (Cutler, 2003; Jescheniak, Hahne, and Schriefers, 2003; Levelt, 2001). Automaticity is required for the speaker/listener to be able to focus on what needs to be said rather than on how the language code is structured or analyzed. By measuring basic encoding and decoding of oral language as performed in integrated tasks in real time, the VAT test probes the degree of automaticity in language performance.

Two basic types of scores are produced from the test: scores relating to the content of what a test-taker says and scores relating to the manner of the test-taker's speaking. This distinction corresponds roughly to Carroll's (1961) description of a knowledge aspect and a control aspect of language performance. In later publications, Carroll (1986) identified the control aspect as automaticity, which occurs when speakers can talk fluently without realizing they are using their knowledge about a language.

Some measures of automaticity can be misconstrued as memory tests. Since some VAT tasks involve repeating long sentences or holding phrases in memory in order to assemble them into reasonable sentences, it may seem that these tasks measure memory instead of language ability, or at least that performance on some tasks may be unduly influenced by general memory performance. Note that every Repeat and every Sentence Build item on the test was presented to a sample of educated native speakers of Arabic and at least 80% of the speakers in that educated native speaker sample responded correctly, or with at most one word missed (deleted or substituted). If memory, as such, were an important component of performance on the VAT tasks, then the native Arabic speakers should show greater performance variation on these items according to the presumed range of individuals' memory spans. Also, if memory capacity (rather than language ability) were a principal component of the variation among people performing these tasks, the test would not correlate so closely with other accepted measures of oral proficiency (see Section 7.3, *Concurrent Validity*).

Note that the VAT probes the psycholinguistic elements of spoken language performance rather than the social, rhetorical and cognitive elements of communication. The reason for this focus is to ensure that test performance relates most closely to the test-taker's facility with the language itself and is not confounded with other factors. The goal is to disentangle familiarity with spoken language from cultural knowledge, understanding of social relations and behavior, and the test-taker's own cognitive style and strengths. Also, by focusing on context-independent material, less time is spent developing a background cognitive schema for the tasks, and more time is spent collecting real performance samples for language assessment.

The VAT test provides a measurement of the real-time encoding and decoding of spoken Arabic. Performance on VAT items predicts a more general spoken Arabic facility, which is essential for successful oral communication in MSA. The same facility in spoken Arabic that enables a person to satisfactorily understand and respond to the listening/speaking tasks in the VAT test also enables that person to participate in native-paced MSA conversation.

## 3. Content Design and Development

### 3.1 Rationale

All VAT item content is designed to be region-neutral. The content specification also requires that both native speakers and proficient non-native speakers find the items easy to understand and to respond to appropriately. For Arabic learners, the items probe a broad range of skill levels and skill profiles.

Except for the Reading items, each VAT item is independent of the other items and presents context independent, spoken material in Arabic. Context-independent material is used in the test items for three reasons. First, context-independent items exercise and measure the most basic meanings of words, phrases, and clauses on which context-dependent meanings are based (Perry, 2001). Second, when language usage is relatively context-independent, task performance depends *less* on factors such as world knowledge and cognitive style and *more* on the test-taker's facility with the language itself. Thus, the test performance relates most closely to language abilities and is not confounded with other test-taker characteristics. Third, context-independent tasks maximize response density; that is, within the time allotted for the test, the test-taker has more time to demonstrate performance in speaking the language because less time is spent presenting contexts that situate a language sample or set up a task demand.

### 3.2 Vocabulary Selection

The items in a Versant-type test are usually limited to include only vocabulary from the most frequently-occurring several thousand words as observed in a substantial corpus of spontaneous speech in the language. However, Arabic morphology is complex. For example, verbs can have prefixes, infixes, and suffixes, and sometimes combinations of these; nouns also have prefixes and suffixes, and prepositions take suffixes. Thus, a count of unique base-form words (as listed in a lexicon) from transcribed spontaneous speech (or even from printed text) requires significant human intervention to produce a lexical frequency table such as is used for English. Available Arabic word lists comprise inflected word forms (not base forms), and in the absence of diacritics and context, Arabic words show much more ambiguity with regard to their base form than English words. Available Arabic word lists are suitable for building a spell checker, but do not offer the information needed for item development.

For this reason, a consolidated word list for Arabic was assembled from the vocabularies of a number of well-known beginning and intermediate Arabic textbooks, along with other instructional materials from the Defense Language Institute (DLI) and from Stanford University. The texts are all listed in Textbook References (Section 11), and include Al-Kitaab, Ahlan wa Sahlan, and the Michigan Modern Standard Arabic books. The general rule for base-form words was: if a word occurs in more than one vocabulary list, then it is included in the consolidated word list. Thus, from the total of 8,419 unique words found in any of the textbook lexicons, the actual consolidated word list that was used to review the draft item texts had 3,296 words. Among the approximately 2,700 unique word stems in the scored item texts, about 80% are contained in the consolidated word list. Note that this does not mean that 20% of the words in running item text that a test-taker encounters are outside this lexicon.

The words from outside the consolidated word list tend to occur relatively infrequently in the item texts themselves. Within a sample of running item text, 93 word tokens in every 100 tokens are from the consolidated word list. Comparison of another sample of item texts with an independent list of common Arabic words (Abdel-Massih's Pan-Arabic lexicon, 1975) indicated that 91% of running item text is contained in that list, which further supports the appropriateness of the item vocabulary.

### 3.3 Item Development

The VAT item texts were drafted by four native speakers of Arabic; all educated in MSA through university level. Two of the item text writers were professors of Arabic teaching at major American universities, and two were adults who had worked professionally in their home countries and had resided in the U.S. for less than two years. In general, the language structures used in the test were designed to reflect those that are common in MSA. In order to make sure that these language structures are indeed used in spoken MSA, many of the language structures were adapted from widely accepted media sources such as Al-Jazeera and AlArabiya.net. Those spoken materials were then altered for appropriate vocabulary and for neutral content. The items were designed to be independent of social and cultural nuance, and high-cognitive functions.

Draft items were then sent for external review to ensure that they conformed to MSA usage in different countries. Dialectically distinct native Arabic-speaking linguists reviewed the items to identify any geographic bias and non-MSA usage. The reviewers were from Egypt, Tunis, Palestine, Syria, and Lebanon – all currently teaching Arabic in American universities. All items, including anticipated responses for short-answer questions, were checked for compliance with the vocabulary specification. Most vocabulary items that were not present in the lexicon were changed to other lexical stems that were in the consolidated word list. Some off-list words were kept and added to a supplementary vocabulary list, as deemed necessary and appropriate. The changes proposed by the different reviewers were then reconciled and the original items were edited accordingly.

### 3.4 Item Prompt Recording

#### 3.4.1 Voice Distribution

Fifteen native speakers (eight men and seven women) representing several different regions were selected for recording the spoken prompt materials. The fifteen speakers recorded the items across different item types fairly evenly.

Recordings were made in a professional recording studio in Menlo Park, California. In addition to the Arabic instruction prompts, English prompts were recorded by an English examiner voice. Thus, there are two versions of the Versant Arabic Test – one with an Arabic examiner voice and another with an English examiner voice.

One specified test development goal was to have the items recorded by voices from a range of locales, including Egypt, Iraq, and the Levant (Lebanon, Syria, Palestine). Table 2 below displays the distribution of item voices across five distinct regions and between genders. Note that the number of items in

Table 2 does not include the 428 Reading items, for which no item-specific recording is played during test administration. The balance of male to female voices in the VAT items is 61% male to 39% female.

Table 2. Distribution of item prompt voices by gender and geographic origin.

Voice	Egypt	Levant	Iraq	Jordan	Morocco	Total
Male	17%	27%	10%	8%	0%	61%
Female	3%	18%	11%	0%	7%	39%
Total	20%	45%	21%	8%	7%	100%

### 3.4.2 Recording Review

The recordings were reviewed for quality, clarity, and conformity to contemporary MSA. In order to evaluate the recordings, a set of error-type taxonomical rubrics were devised by an Arabic linguist. Any recording in which reviewers noted non-phonological deviations from MSA was excluded from installation in the operational test.

## 4. Score Reporting

### 4.1 Scores and Weights

Of the 69 items in an administration of the Versant Arabic Test, 62 responses are used in the automatic scoring. The first item response of each task type in the test is considered a practice item and is not incorporated into the final score. In addition, the three passage retelling responses in Part F are not scored automatically.

The VAT score report is comprised of an Overall score and four diagnostic subscores (Sentence Mastery, Vocabulary, Fluency<sup>1</sup>, and Pronunciation).

**Overall:** The Overall score of the test represents the ability to understand spoken MSA and speak it intelligibly at a native-like conversational pace on common topics. Overall scores are based on a weighted combination of the four diagnostic subscores (30% Sentence Mastery, 20% Vocabulary, 30% Fluency and 20% Pronunciation). All scores are reported in the range from 20 to 80.

**Sentence Mastery:** Sentence Mastery reflects the ability to understand, recall, and produce Arabic phrases and clauses in complete sentences. Performance depends on accurate syntactic

<sup>1</sup> Within the context of language acquisition, the term “fluency” is sometimes used in the broader sense of general language mastery. In the narrower sense used in VAT score reporting, “fluency” is taken as a component of oral proficiency that describes certain characteristics of the observable performance. Following this usage, Lennon (1990) identified fluency as “an impression on the listener’s part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently” (p. 391). In Lennon’s view, surface fluency is an indication of a fluent process of encoding. The VAT fluency subscore is based on measurements of surface features such as the response latency, speaking rate, and continuity in speech flow, but as a constituent of the Overall score it is also an indication of the ease of the underlying encoding process.



processing and appropriate usage of words, phrases, and clauses in meaningful sentence structures.

**Vocabulary:** Vocabulary reflects the ability to understand common words spoken in sentence context and to produce such words as needed. Performance depends on familiarity with the form and meaning of common words and their use in connected speech.

**Fluency:** Fluency is measured from the rhythm, phrasing and timing evident in constructing, reading and repeating sentences.

**Pronunciation:** Pronunciation reflects the ability to produce consonants, vowels, and stress in a native-like manner in sentence context. Performance depends on knowledge of the phonological structure of common words.

Figure 2 illustrates which sections of the test contribute to each of the four subscores. Each vertical rectangle represents the response utterance from a test-taker. The items that are not included in the automatic scoring are shown in blue. These include the first item in each of the first four sections of the test and all three items in the last section (Passage Retellings).

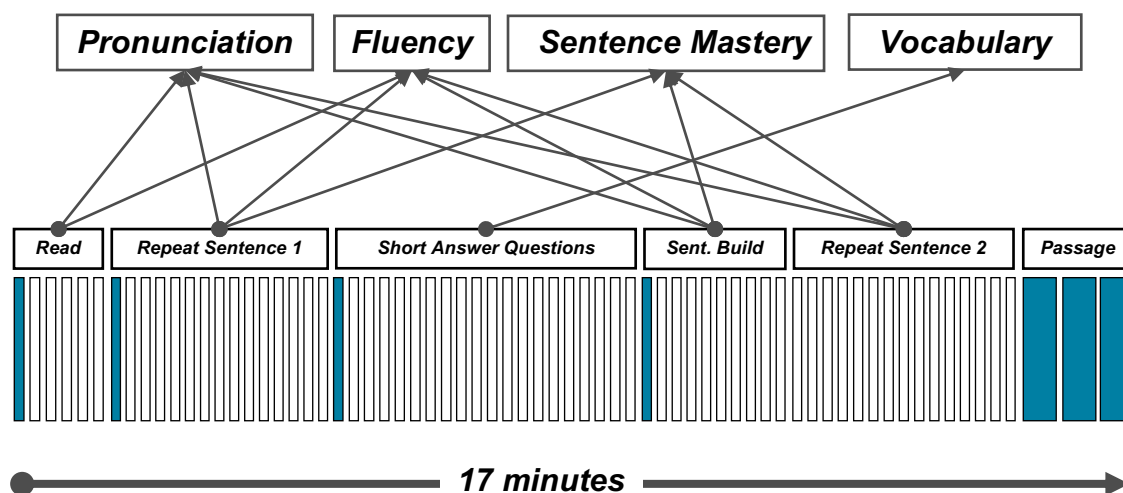


Figure 2. Relation of subscores to item types.

The subscores are based on two different aspects of language performance: a knowledge aspect (the content of what is said), and a control aspect (the manner in which a response is said). The four subscores reflect these aspects of communication where Sentence Mastery and Vocabulary are associated with content and Fluency and Pronunciation are associated with manner of speaking. The content accuracy dimension counts for 50% of the Overall score and indicates whether or not the test-taker understood the prompt and responded with appropriate content. The manner-of-speaking scores count for the remaining 50% of the Overall score, and indicate whether or not the test-taker speaks like an educated native (or like a very high-proficiency learner). Producing accurate lexical and structural content is important, but excessive attention to accuracy can lead to disfluent speech.



production and can also hinder oral communication; on the other hand, inappropriate word usage and misapplied syntactic structures can also hinder communication. In the VAT scoring logic, content and manner (i.e., accuracy and control) are weighted equally because successful communication depends on both.

The Ordinate automated scoring system scores both the content and manner-of-speaking subscores using a speech recognition system that is optimized based on non-native Arabic spoken response data collected during the field test. The content subscores are derived from the correctness of the test-taker's response and the presence or absence of expected words in correct sequences. The manner-of-speaking subscores (Fluency and Pronunciation, as the control dimension) are calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. In order to produce valid scores, during the development stage, these measures were automatically generated on a sample set of utterances (from both native and non-native speakers) and then were scaled to match human ratings. Three trained native speakers rated test-takers' pronunciation and fluency from the Reading, Repeat, and Sentence Build speech files, with reference to a set of rating criteria. These criterion-referenced human scores were then used to rescale the machine scores so that the pronunciation and fluency subscores generated by the machine align with the human ratings.

## 4.2 Score Use

Once a test-taker has completed a test, the Ordinate testing system analyzes the spoken performances and posts the scores at [www.pearson.com/english](http://www.pearson.com/english). Test administrators and score users can then view and print out the test results from the Versant website, as well as listen to selected responses from each test-taker.

Score users of the Versant Arabic Test may be educational and government institutions or commercial and business organizations. Within a pedagogical research setting, VAT scores may be used to evaluate the level of spoken Arabic skills of individuals entering into, progressing through, and leaving Arabic language courses.

# Section II – Field Test and Validation Studies

## 5. Field Test

### 5.1 Data Collection

Both native speakers and non-native speakers of Arabic were recruited as participants from October 2007 through August 2008 to take a prototype data-collection version of the Versant Arabic Test. The purposes of this field testing were 1) to validate operation of the test items with both native and non-native speakers, 2) to calibrate the difficulty of each item based on a large sample of test-takers at various levels, and 3) to collect sufficient Arabic speech samples to train and optimize the automatic

speech processing system, and to develop automatic scoring models for spoken Modern Standard Arabic.

### 5.1.1 Native Speakers

A total of 1,373 adult native Arabic speakers were recruited from different Arabic-speaking countries. These 1,373 speakers produced a total of 1615 completed tests, as some native speakers took the test more than once. Native speakers are defined as individuals who spent the first fifteen years of their lives in an Arabic-speaking country and were educated in Arabic through the university level. Samples were gender balanced when possible.

After the data collection, two educated native Arabic speakers listened to a sample of responses from the tests and evaluated the quality of the tests – both for signal quality, and to identify any of the native Arabic speakers who were obviously not educated in MSA. Any putative native test with either very poor signal quality or evidence of poor control of MSA forms was then excluded from the development process. This review process yielded a vetted native sample of 1,316 completed tests. The native sample includes speakers from 18 different countries, including at least 30 speakers from each of Egypt, Syria, Palestine, Iraq, Sudan, Yemen, Morocco, Jordan, and Algeria. The male-to-female ratio was 57-to-43. The average age among the native test-takers was 29.5 years old.

### 5.1.2 Non-Native Speakers

Most of the test-takers in the non-native speaker sample were students at the Defense Language Institute Foreign Language Center (DLI-FLC) in Monterey, California. A total of 735 students of Arabic at the DLI-FLC participated, with some students taking the test twice. An additional 552 tests were completed by learners of Arabic not affiliated with DLI. The dominant first language of the non-native sample was English. Other major first languages included Kurdish (45 subjects) and Somali (28 subjects). Among the non-natives, 31 subjects reported being heritage speakers of Arabic. The average age of the learner sample was 25.2 years old, with 65% of the sample male and 35% female. A total of 1,332 learner tests were used in test development. The mean VAT Overall score of this sample was 44.2 with a standard deviation of 16.6. The standard error of the VAT Overall score is 2.2 points on the 60-point reporting scale.

## 6. Data Resources for Score Development

### 6.1 Data Preparation

During the development of the VAT, a total of 264,000 spoken responses were collected from natives and learners. The vast majority of the native responses were transcribed at least once, and almost all the non-native responses were transcribed two or more times. Subsets of the response data (from outside the set of validation test-takers) were also presented to native listeners for quality and/or proficiency judgments so they could be used in score calibration.

#### 6.1.1 Transcription

Both native and non-native responses were transcribed by native speakers of Arabic in order to train an automatic speech recognition system that is optimized for non-native speech patterns. A total of 126,535 transcriptions were produced for native responses and 259,424 transcriptions were produced for non-native responses. The native speaker transcribers were rigorously trained and the quality of their transcriptions was closely monitored.

### 6.1.2 Human Rating

Selected item responses from a subset of test-takers were presented to three educated native Arabic speakers to be judged for pronunciation, fluency, and as indicators of overall proficiency. Before the native speakers began rating responses, they received training in how to evaluate responses according to analytical and holistic rating criteria. Each rater listened to the sets of item response recordings in a different random order, and independently assigned pronunciation, fluency, and proficiency scores. The raters called in to the Ordinate system and were first presented with a set of responses for the pronunciation rating, and then were presented with a different (but slightly overlapping) set of responses for the fluency rating, and finally were presented with another set for proficiency judgments. Separating the judgment of different traits is intended to minimize the transfer of judgments from pronunciation to fluency or to overall proficiency, by having the raters focus on only one trait at a time. For the fluency and pronunciation sets, rating stopped when each item had been judged by each of the three raters; with the proficiency sets, the group of three raters produced two judgments total per item.

## 7. Validation

### 7.1 Validity Study Design

Validity analyses examined three aspects of the VAT scores:

1. Internal quality (reliability and accuracy): whether or not the VAT provides consistent scores that accurately reflect the scores that human listeners and raters would assign.
2. Relation to known populations: whether or not the VAT scores reflect expected differences and similarities among known populations (e.g., natives vs. learners).
3. Relation to scores of tests with related constructs: how closely do VAT scores predict the reliable information in scores of well-established speaking tests (e.g., the Interagency Language Roundtable Oral Proficiency Interview (ILR OPI)).

From the large body of spoken performance data collected from native speakers of Arabic and learners of Arabic during the field studies, qualified data was analyzed to provide evidence relevant to the validity of VAT as a measure of proficiency in spoken Arabic. A sample of test-takers was kept separate from the development data, so that the sample could be used as an independent validation set, as shown in Figure 3.

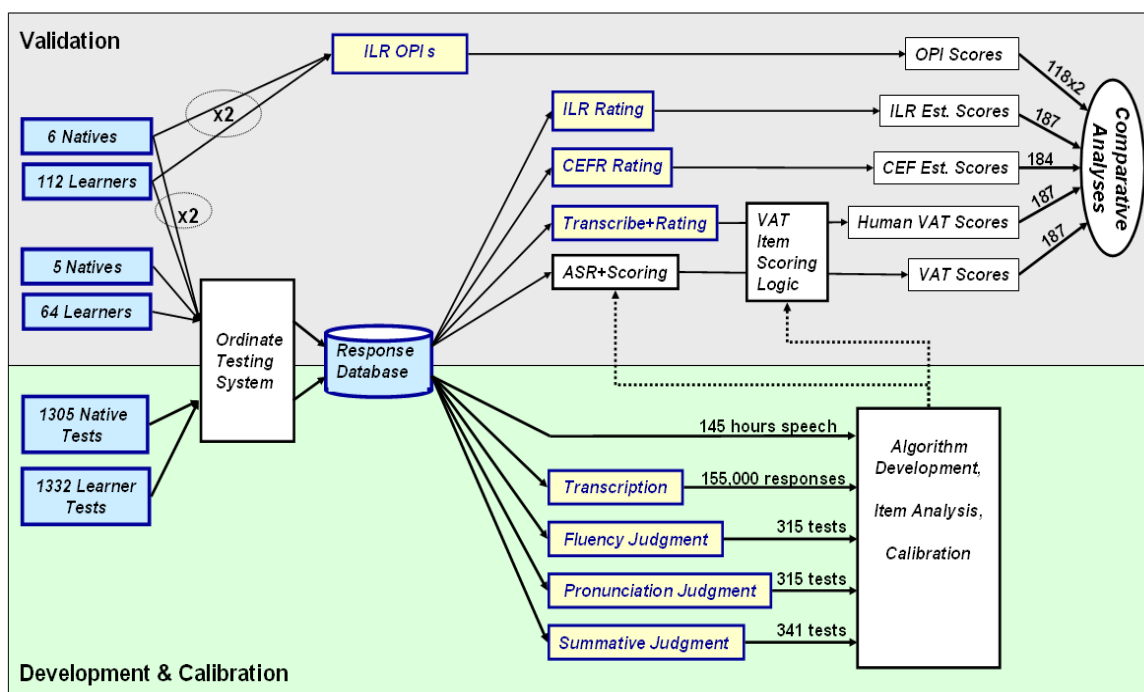


Figure 3. Data flow and separation during VAT development and validation

### 7.1.1 Validation Sample

A total of 187 subjects were set aside for a series of validation analyses. Care was taken to ensure that the training dataset and validation dataset did not overlap. The spoken performance samples provided by the validation test-takers were excluded from the datasets used for training the automatic speech processing models or for training any of the scoring models. Of these 187 subjects, 118 also took the ILR OPI for concurrent validation. All of these concurrent OPI subjects (N = 118) were included only in this validation sample and excluded from the training dataset. The remaining 69 subjects were selected in order to produce a relatively flat distribution of proficiency levels in the validation sample. The selection was informed by scores from overall proficiency ratings provided by two Arabic native speakers. A total of 11 native speakers were also included in this sample – six in the OPI group and five among the extra subjects. Many of the internal validation indices (e.g. split-half reliability and standard error of measurement (SEM)) were measured on a sub-sample of 134 test-takers from the validation sample. This sub-sample includes the same six native speakers found in the set of 118 test-takers who completed OPI tests.

### 7.1.2 Test Materials

Note that the data-collection test administered in the validation studies was longer than the VAT described in this document. The data-collection test had a total of 45 Repeat items presented in three Repeat sections instead of a total of 30 Repeat items in two Repeat sections. The data-collection tests were presented with either 80 or 81 items, instead of the 69 items in the VAT test product. From each data-collection test, a virtual VAT was reconstituted with the final VAT item counts and used in the analysis reported here.

## 7.2 Internal Validity

### 7.2.1 Validation Sample Statistics

A total of 1,332 learner tests were collected for use in test development. The mean Overall score of this sample was 44.15 with a standard deviation of 16.62. Table 3 summarizes some descriptive statistics for the non-native sample used for the development purpose and for the validation sample. The mean score of the validation sample is higher because the 11 native speakers are included in this validation set, as described above.

Table 3. Overall score statistics for non-native development sample and validation sample.

Overall Score Statistics	Non-Native Development Sample (N=1332)	Validation Sample (N=187)
Mean	44.15	55.89
Standard Error	0.46	1.57
Median	41.66	57.75
Sample Variance	276.11	458.11
Kurtosis	0.34	-1.22
Skewness	0.82	-0.12

### 7.2.2 Test Reliability

To understand the consistency and accuracy of VAT Overall scores and the distinctness of the subscores, the following indicators were examined: the SEM of the VAT Overall score, the reliability of the VAT (split-half and test-retest reliability); the correlations between the VAT Overall score and the VAT subscores, and between pairs of VAT subscores; comparison of machine-generated VAT scores with listener-judge scores of the same VAT tests. These qualities of consistency and accuracy of the test scores are the foundation of any valid test (Bachman & Palmer, 1996).

The SEM provides an estimate of the amount of error in an individual's observed test score and "shows how far it is worth taking the reported score at face value" (Luoma, 2003: 183). The SEM of the VAT Overall score is 2.2.

Score reliabilities were estimated by both the split-half method ( $n = 134$ ) and also by the test-retest method ( $N = 100$ ). Split-half reliability was calculated for the Overall score and all subscores. The split-half reliabilities use the Spearman-Brown Prophecy Formula to correct for underestimation and they

are similar to the reliabilities calculated for the uncorrected test-retest data set (see Table 4). All test-retest subjects took the two VAT tests within a 15-day window.

In both split-half and test-retest estimations, the reliability for the Vocabulary subscore is lower than the reliability of the other subscores. This may be because the score is based on a small number of items that do not sample the target vocabulary widely enough. That is, variability in test-takers' vocabulary knowledge across test items is large in relation to the small sample of vocabulary presented in the Short Answer items that are used as the sole basis for the Vocabulary subscore.

Table 4. Reliability of Versant Arabic Test machine scoring.

Score	Split-half reliability (N = 134)	Test-retest reliability (nN = 100)
<b>Overall</b>	0.98	0.97
<b>Sentence Mastery</b>	0.97	0.96
<b>Vocabulary</b>	0.89	0.82
<b>Fluency</b>	0.97	0.96
<b>Pronunciation</b>	0.96	0.94

### 7.2.3 Dimensionality: Correlations between Subscores

Table 5 presents the correlations between pairs of VAT subscores and between the subscores and the Overall score for the validation sub-sample of 134 test-takers, which includes six native speakers.

Table 5. Correlations between VAT subscores for the validation sub-sample (N = 134).

Correlation	Vocabulary	Pronunciation	Fluency	VAT Overall
<b>Sentence Mastery</b>	0.89	0.79	0.75	<b>0.95</b>
<b>Vocabulary</b>		0.74	0.71	<b>0.92</b>
<b>Pronunciation</b>			0.94	<b>0.92</b>
<b>Fluency</b>				<b>0.90</b>

Test subscores are expected to correlate with each other to some extent by virtue of presumed general covariance within the test-taker population between different component elements of spoken language skills. However, the correlations between the subscores are significantly below unity, which indicates that the different scores measure different aspects of the test construct. These results reflect the operation of the scoring system as it applies different measurement methods and uses different aspects of the responses in extracting the various subscores from partially overlapping response sets.

Figure 4 illustrates the relationship between two relatively independent machine scores (Sentence Mastery and Fluency) for only the learners of Arabic the people in the validation sample (N = 128) who did the OPI tests. These machine scores are calculated from a subset of responses that are mostly overlapping (Repeats and Sentence Builds for Sentence Mastery and Repeats, Sentence Builds and Readings for Fluency). Although these measures are derived from overlapping sets of responses, the

subscores clearly extract distinct measures from these responses. For example, test-takers with Fluency scores in the 30-50 range have Sentence Mastery scores that are spread roughly evenly over the whole 20-80 score range.

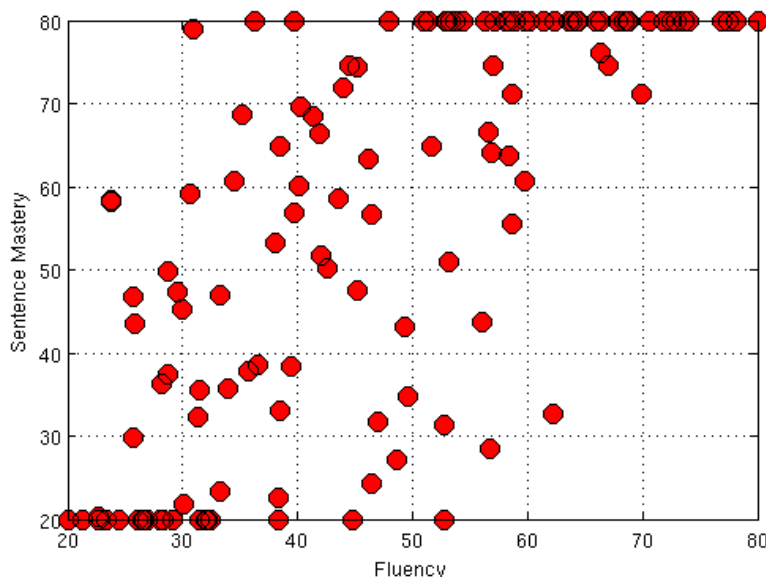


Figure 4. Fluency versus Sentence Mastery scores for the learner-only subset of the OPI takers in the validation sample ( $N = 128$ ;  $r = 0.71$ )

#### 7.2.4 Machine Accuracy: VAT Scored by Machine vs. Scored by Human Raters

Table 6 presents Pearson product-moment correlations between machine scores and human scores, when both methods are applied to the same performances on the same VAT responses. The test-taker set is the same sub-sample of 134 test-takers that was used in the reliability and subscore analyses. Correlations presented in Table 5 suggest that scoring a Versant Arabic Test by machine will yield scores that generally correspond with human scoring of the same material. Among the subscores, the human-machine relation is closer for the content accuracy scores than for the manner-of-speaking scores, but the relation is close for all four subscores. At the Overall score level, Versant Arabic Test machine-generated scores are virtually indistinguishable from scoring that is done by careful human transcriptions and multiple independent human judgments. That is, at the Overall score level, machine scoring does not introduce any substantial error in the scoring.

Table 6. Correlations between human and machine scoring of the Versant Arabic Test ( $N = 134$ ).

Score	Correlation
Overall	0.97
Sentence Mastery	0.97
Vocabulary	0.96
Fluency	0.84
Pronunciation	0.83



### 7.2.5 Differences among Known Populations

Examining how different groups perform on the test is another aspect of test score validity. In most language tests, one would expect most educated native speakers to get good scores, and for learners, as a group, to perform less well. Also, if the test is designed to distinguish ability levels among learners, the scores should be distributed widely over the learners. Because MSA is an educated, prestige form of the Arabic language, one expects that there may be some notable difference in VAT score distributions between groups of educated and uneducated native speakers of Arabic. However, one might expect that there are no important differences among national groups of educated native speakers. Educated Arabs, whether from Palestine, or Egypt, or Saudi Arabia, for example, should do equally well on the test.

Group Performance: Figure 5 presents cumulative distribution functions for two speaker groups: educated native speakers of Arabic and learners of Arabic. The Overall score and the subscores are reported in the range from 20 to 80 (scores above 80 are reported as 80 and scores below 20 are reported as 20). For this reason, in Figures 5 and 6, the trajectory of the Overall score curve above 80 and below 20 is shown in a shaded area. One can see that the distribution of the native speakers clearly distinguishes the natives from the learner sample. For example, fewer than 5% of the native speakers score below 70, while fewer than 10% of the learners score above 70.

The distribution of learner scores suggests that the Versant Arabic Test has high discriminatory power among learners of Arabic as a second or foreign language, whereas educated native speakers obtain maximum or near-maximum scores. Note that natives in Figure 5 are all nominally graduates from Arabic-language universities.

Because MSA is a prestige form of Arabic that is usually learned in school, it should be instructive to compare the test performance of educated and uneducated native speakers of Arabic. To this end, a small convenience sample of 29 Egyptians working in non-professional positions in Cairo was recruited to take VAT. The cumulative distribution of their 29 Overall scores in relation to the distribution of the 1,309 educated natives is shown in Figure 6.



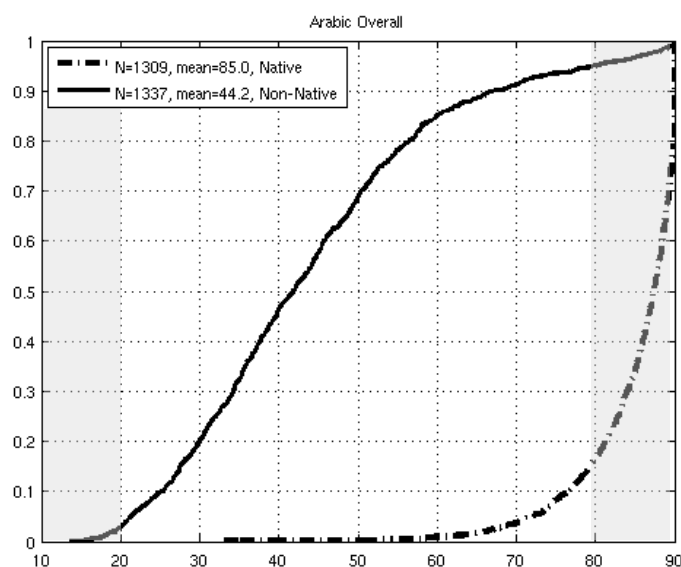


Figure 5. Cumulative distributions for educated native speakers and learners of Arabic.

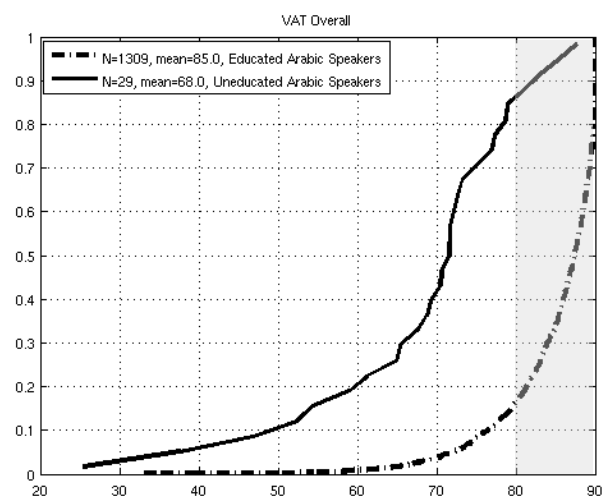


Figure 6. Cumulative distributions for educated and uneducated native speakers of Arabic.

Although there is a great difference in sample size between the two groups shown in Figure 6, some differences are apparent. The median Overall score of the uneducated native speakers is 72 whereas the median score for the educated natives is 87 (reported as 80). The cumulative functions seem notably different, with only 14% of the uneducated speaker group receiving the highest Overall score (80). Although native skill in a colloquial form of Arabic does not by itself match the MSA skills of the educated natives, if one compares the uneducated natives with the cumulative curve of learners in Figure 5, it is evident that VAT Overall scores would indicate that almost all uneducated native speakers do much better with MSA than most learners do.

An analysis of variance (ANOVA) was conducted to verify that the Versant Arabic Test was not sensitive to differences in subject characteristics that are not relevant to the construct. Unrelated speaker characteristics include national origin and speaker gender.

Since the test was normalized with a preponderance of Arabic speakers from Egypt and Syria, a further validation experiment was conducted to investigate whether or not speakers from other Arab countries perform just as well as the Egyptians and Syrians. The cumulative distribution functions of educated native Arabic test-takers from seven different countries are presented in Figure 7. As seen in Figure 7, the Iraqi, Palestinian, Saudi and other speakers perform just as well as speakers from Egypt and Syria. The findings support the notion that the Versant Arabic Test scores reflect a speaker's facility in spoken MSA, irrespective of the speaker's country of origin.

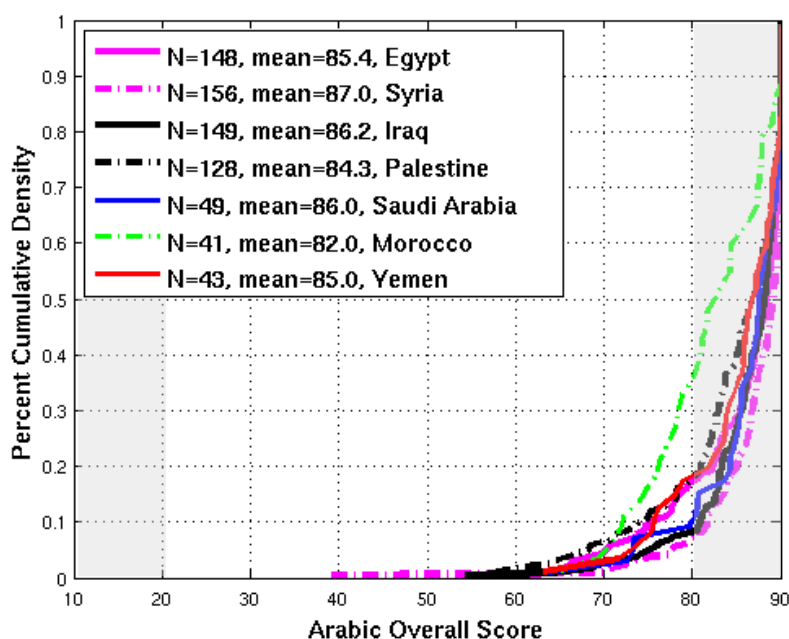


Figure 7. Cumulative distribution functions for native Arabic speakers of different countries.

In order to check if the VAT scores are sensitive to test-taker gender, a subset of the validation sample (134 test-takers; 58 women and 76 men) was subjected to a 2x2 mixed factor ANOVA, with score method (human or machine) as the within-subjects factor, and gender (male or female) as the between-subjects factor. No significant main effect of scoring method or test-taker gender was found, nor was the interaction significant (all  $p$ 's > 0.10).

## 7.3 Concurrent Validity: Correlations between VAT and Human Scores

### 7.3.1 Concurrent Measures

One important goal of the validity studies is to understand how the automated VAT scores relate to other measures of spoken Arabic that are scored by trained human raters. Human-administered, human-rated oral proficiency interviews are widely accepted as the "gold standard" measure of spoken

communication skill, or oral proficiency. The ILR-OPI is a human assessment method that is of particular interest because its scores are used widely within the U.S. government. For this reason, automatically derived VAT Overall scores were compared to two types of human-generated scores. The four measures used in this concurrent validity study are:

- |                          |  |
|--------------------------|--|
| (1) VAT Overall scores   | {20 – 80}                              |
| (2) ILR OPI scores       | {0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5} |
| (3) ILR Level Estimates  | {0, 0+, 1, 1+, 2, 2+, 3, 3+, 4, 4+, 5} |
| (4) CEFR Level Estimates | {A1-, A1, A2, B1, B2, C1, C2}          |

The VAT Overall score (1) is the weighted average of four subscores automatically derived from 62 spoken responses as schematized in Figure 2. The ILR OPI scores (2) are an amalgam of four reported ILR scores per test-taker (two scores from each of two interviews). The ILR Level Estimates (3) are an amalgam of multiple estimates of test-taker level based on samples of recorded responses from the unscored Passage Retellings for each test-taker. The Common European Framework of Reference (CEFR; Council of Europe, 2003) Level Estimates (4) are an amalgam of multiple estimates of test-taker level based on samples of recorded responses from the unscored Passage Retellings for each test-taker.

Method of ILR OPI ratings: During field testing of the VAT, a sample of 112 non-native test-takers and six Arabic native test-takers were asked to concurrently take two Versant Arabic tests and two OPIs following the protocol of the Interagency Language Roundtable (ILR-OPIs). Of the non-native speakers in this concurrent testing sample, at least 20 test-takers were learning Arabic at a college in the U.S., at least 11 were graduates from the Center for Arabic Studies Abroad program, nine test-takers were recruited at a private language school in Cairo, Egypt, and the remainder were current or former students of the Arabic language who were recruited in the U.S. The mean age of the test-takers was 27 years old ( $SD = 7$ ) and the male-to-female ratio was 60 to 58. The first language (L1) of 93 test-takers was English, followed by six native Arabic speakers, four Danish speakers, three Arabic speakers who grew up outside of the Middle East, two Spanish speakers, two Somali speakers, one Aramaic, one Persian, one German, one Hebrew, one Italian, one Japanese, one Polish, and one Urdu speaker. 17 test-takers reported that they were heritage Arabic speakers. They all took both OPIs via telephone, and 114 out of the 118 participants also took both Versant tests via telephone (the remaining four participants took both of their Versants tests via computer). Of the 118 subjects, 114 completed all four of the tests within a ten-day window, and the remaining four subjects within a 15-day window.

Seven active government-certified oral proficiency interviewer-raters conducted the ILR OPIs over the telephone, following the official ILR procedures. Six of the seven raters were grouped into three pairs and each pair conducted an OPI with each test-taker independently. The seventh interviewer-rater served as a fill-in for the other three pairs when one interviewer could not be present. The interviewers were instructed to independently submit their own ILR level rating directly to Pearson via e-mail, without any discussion with the other interviewer/rater in the pair.

Method of ILR Level Estimates: Two or more weeks after these concurrent tests, the same OPI raters listened independently (and in different random orders) to two VAT passage retelling responses from

each of the 118 test-takers and assigned the test-taker's estimated ILR level to each of the passage retelling responses. These ratings are referred to as *ILR Level Estimates* in the text below.

**Method of CEFR Level Estimates:** An experienced language testing expert trained four educated native speakers of Arabic in using the CEFR scale. After the familiarization/training session, the four raters independently listened to a set of VAT passage retelling responses from each of the 187 test-takers and assigned an estimated CEFR level to each of the test-takers. These ratings are referred to as *CEFR Level Estimates*.

### 7.3.2 OPI Reliability

The reliability of the OPI test scores was calculated by assigning numerical values to the scores, with a "+" score taken as +0.5. For example, a reported score of 2+ was thus taken to have the value 2.5. The interview score for a particular interview was taken to be the simple average of the two independent scores assigned by the two interviewer/raters. For the set of 118 test-takers in the validation sample, the Pearson product-moment correlation between the two interview scores was 0.91. This indicates that the independent interview scores are consistent.

The predictive power of the seven individual raters can be seen by comparing each rater's individual scores with the average score from the other interview. In Table 7, the correlation coefficients range from 0.86 to 0.93, suggesting that, even across interviews with the same test-taker, all the individual interviewer/raters were quite consistent. The average inter-rater reliability, using Fisher's Z transformation, is 0.90.

Table 7. Correlations between individual interviewer's ILR ratings and the combined ratings of two other interviewers from a different interview with the same subjects.

Interviewer/Rater	N	Correlation
A	85	0.89
B	84	0.86
C	92	0.90
D	76	0.91
E	49	0.93
F	53	0.92
G	33	0.86
<b>Average</b>		<b>0.90</b>

Together, these data indicate that the selected human raters were consistent among the group, and the ILR-OPI procedure as implemented here was consistent. The test-retest reliability of the VAT is 0.97 (see Table 3 above), indicating that the VAT was also reliable across administrations.

### 7.3.3 VAT and ILR OPIs

For each of the test-takers, the VAT overall score was correlated with a criterion ILR level score that was calculated from all the available independent ILR rater judgments. The criterion ILR score was derived from the ordinal ILR ratings using a Rasch model as implemented in the computer program FACETS (Linacre, 2003). The FACETS program estimates rater severity, subject ability, and item difficulty (Linacre, Wright, and Lunz, 1990). The model assumes a single underlying dimension, where values along this dimension are expression of the ability of test-takers in logits. For all the human ratings (ILR OPIs or ILR Estimates) displayed in figures below, the boundaries of the different levels were mapped onto a continuous Logit scale. Figure 8 is a scatter plot of the ILR OPI ability estimates as a function of VAT scores for 118 test-takers.

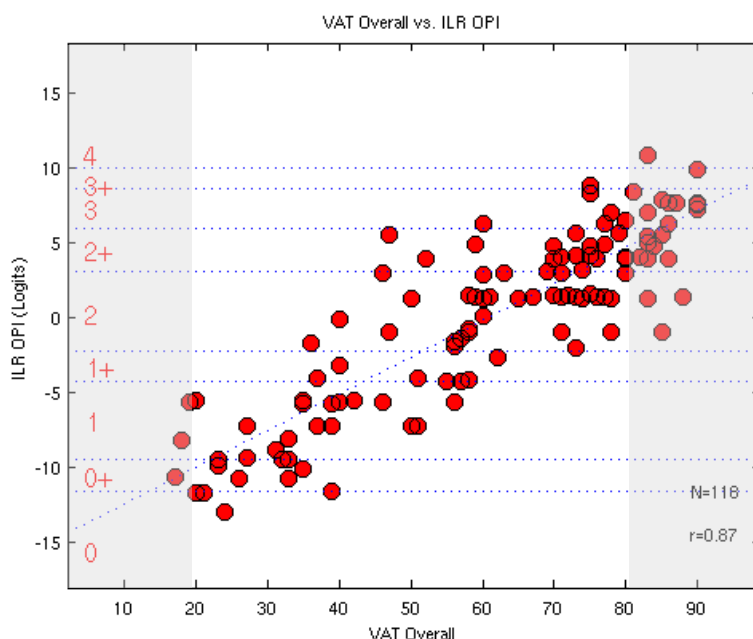


Figure 8. Test-takers' ILR OPI scores as a function of VAT scores (N = 118;  $r = 0.87$ ).

The correlation between these two sets of test scores is 0.87, indicating a reasonably close relation between machine-generated scores and human-rated interviews, in spite of the relative brevity of the VAT (17 minutes vs. 40-60 minutes for the ILR OPI). Note that this correlation of 0.87 is in the same range as those observed between a single certified ILR OPI interviewer and the average of two others conducting an independent interview.

### 7.3.4 VAT and ILR Level Estimates

An additional validation analysis involved comparing the VAT scores with expert human ratings of spoken responses to VAT items. The OPI interviewer/raters listened to and judged VAT response in two sessions. The judgments were estimates of the test-takers' ILR levels extrapolated from the recorded response material. In the first session, four of the OPI raters judged two unscored Passage Retelling responses from each of 53 test-takers during a several-month hiatus in OPI testing. In the second session, several months later, all seven OPI raters listened to and judged two Passage Retelling

responses from each of 134 tests. These judgments were combined using IRT to produce an estimated ILR level for each of the 187 test-takers in the validation sample. These combined scores are referred to as ILR Level Estimates. The sample included 17 Arabic speakers (11 native speakers and 6 heritage speakers).

OPI Ratings and ILR Level Estimates: To evaluate the accuracy of the ILR Level Estimate, the relationship between the full OPIs and the ILR Level Estimates was investigated with the dataset of the 118 speakers who participated in the OPI concurrent study and whose spoken responses were rated by the OPI raters in the ILR Level Estimate study. The full-scale OPI ratings and ILR-Level Estimates follow a roughly linear relationship with a correlation of 0.86, as displayed in Figure 9.

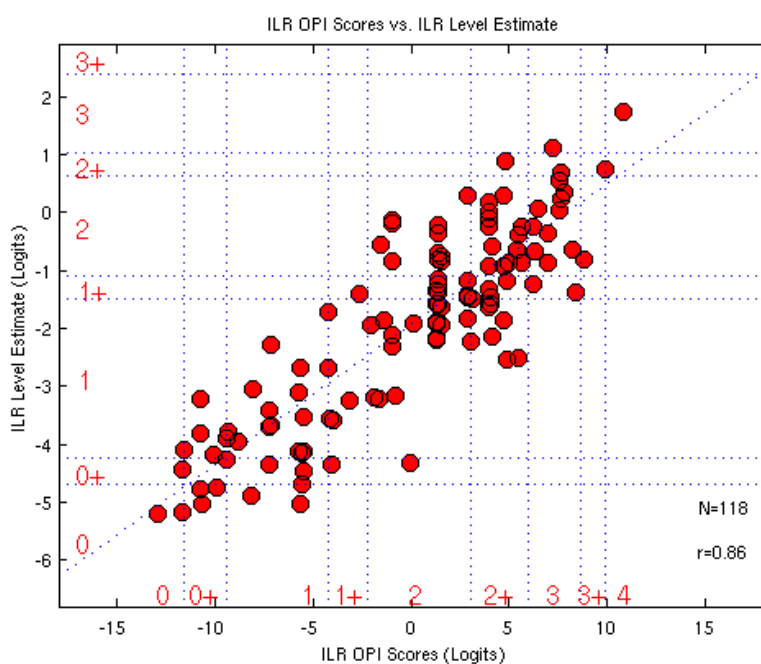


Figure 9. ILR Estimates on spontaneous speech samples as a function of full-length OPI ratings (N = 118;  $r = 0.86$ )

This result suggests that Passage Retelling responses may contain suitable material from which OPI raters can make a reasonable estimate of a speaker's oral proficiency in Arabic. Further, it supports the use of scores from this separate performance to check the VAT scores.

ILR Level Estimates and VAT Scores: When the VAT scores were compared with the ILR Level Estimates, as shown in Figure 10, the correlation coefficient was 0.88. Note that this data set included five native speakers of Arabic who were born and raised in the Middle East. Four other speakers indicated that their first language is Arabic, although they were raised in Canada or Denmark. The rest of the sample was mostly comprised of native English speakers.

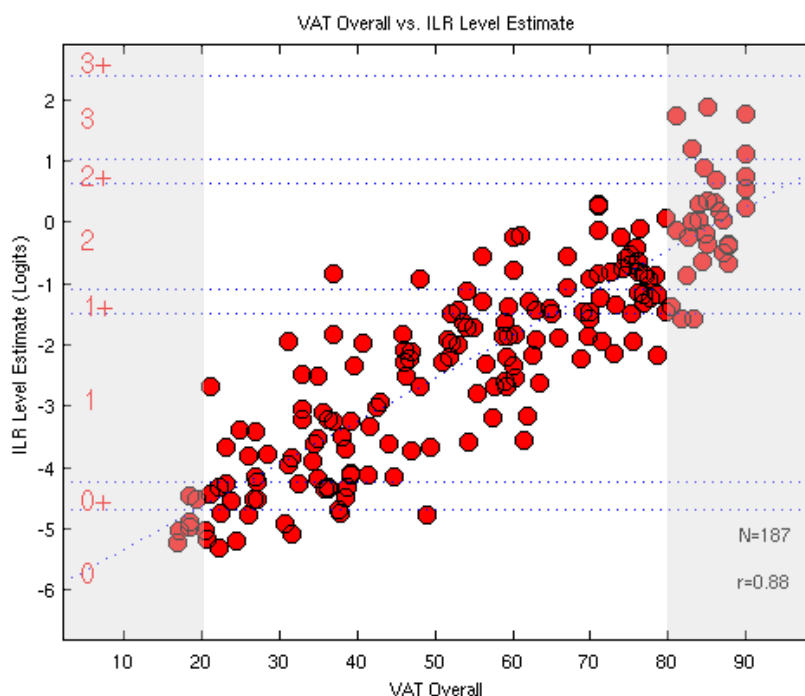


Figure 10. ILR Estimates from speech samples as a function of VAT scores (N = 187;  $r = 0.88$ ).

It is clear from this data that the VAT overall score correlates substantially with the ILR Level Estimates. This can be seen as evidence that the VAT provides useful information related to how well a learner can speak Arabic in the semi-spontaneous task situation of passage retelling.

### 7.3.5 VAT and CEFR Level Estimates

Four educated native Arabic speakers judged a sample of responses of 187 test-takers and provided an estimated level on the CEFR scale (Council of Europe, 2003). The CEFR descriptors used in this validation study were 'Global Oral Assessment Scale' (2003: 24) and 'Oral Assessment Criteria Grid' (2003: 25).

A CEFR familiarization/standardization session was held by a language testing expert who had extensive experience using the CEFR scale. In accordance with recommended procedure, four educated native speakers were trained as raters by undergoing a familiarization session in which they read the scales and listened to benchmark oral performances provided on the Council of Europe website ([www.coe.int/portfolio](http://www.coe.int/portfolio)). The raters then listened to a training set of 20 test-taker performances and decided which CEFR level best described the test-takers. Following this training, each rater independently assessed the performances of all 187 test-takers in the validation set. The performances consisted of test-takers' responses to five passage retell items where the passage items increased gradually in length and complexity from 19 words to 50 words. The raters were encouraged to listen to all five responses from the test-taker and get as full a picture of their proficiency as possible before assigning a CEFR level to that test-taker.

Analysis: The agreement between raters was calculated using a Pearson product-moment correlation for each pair of raters. The average correlation was 0.80, demonstrating sufficient consistency among the raters.

Table 8. Correlations between pairs of Arabic CEFR raters.

Rater Pair	N	Correlation
F1, M1	187	0.83
F1, M2	187	0.79
F1, F2	187	0.78
M1, M2	187	0.79
M1, F2	187	0.81
M2, F2	187	0.78
<b>Average</b>	<b>187</b>	<b>0.80</b>

A multi-facet Rasch model (where the two facets were rater and test-taker) was generated to transform the raters' categorical CEFR levels into continuous logit scores. During the analysis, it was noticed that three test-takers did not answer any of the passage retelling responses, making it impossible for the raters to provide any estimated CEFR level. These three subjects were therefore removed from further analysis. For the remaining 184 subjects, the correlation between the VAT overall scores and logit-based CEFR estimate measures was 0.85, as shown below in Figure 11, suggesting a reasonably strong relationship between VAT overall scores and CEFR levels.

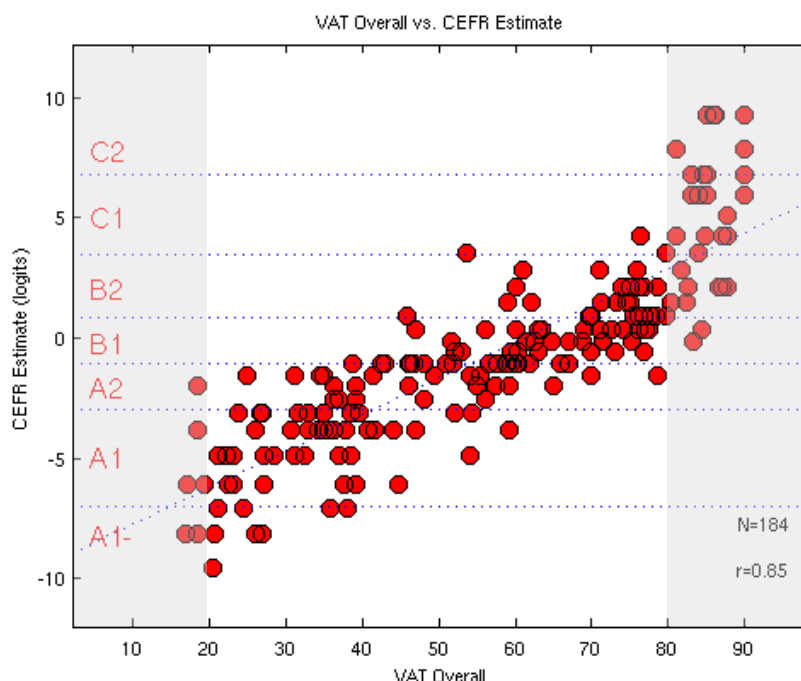


Figure 11. VAT Overall scores versus CEFR estimates (N = 184;  $r = 0.85$ ).



Table 9 below summarizes the correlation coefficients between different measures used in the concurrent validation studies. The range of correlation coefficients between the VAT and the other assessments of oral proficiency is from 0.85 to 0.88. In addition, all the correlations between the other measures of oral proficiency are also highly correlated with each other. Thus, not only does the VAT produce results that are similar to human raters, the VAT test also elicits sufficient spoken language behavior on which to base a reliable human judgment of speaking skills in MSA.

*Table 9. Correlations between VAT and concurrent tests and CEFR Level Estimates.*

Correlation (N)	VAT	ILR OPI	ILR Level Estimate	CEFR Level Estimate
<b>VAT</b>	0.97 (100) <sup>2</sup>			
<b>ILR OPI</b>	0.87 (118)	0.91 (118) <sup>1</sup>		
<b>ILR Level Estimate</b>	0.88 (187)	0.86 (118)	0.90 (187) <sup>2</sup>	
<b>CEFR Level Estimate</b>	0.85 (184)	0.82 (118)	0.88 (184)	0.80 (187) <sup>2</sup>

## 8. Conclusion

This report has provided validity evidence so that score users can make an informed interpretive judgment as to whether or not Versant Arabic Test scores would be valid for their purposes. The test development process is documented and adheres to sound theoretical principles and test development ethics from the field of applied linguistics and language testing, namely: the items were written to specifications and subject to a rigorous procedure of qualitative review and psychometric analysis before being deployed to the item pool; the content was selected from both pedagogic and authentic material; the test has a well-defined construct that is represented in the cognitive demands of the tasks; the scores, item weights and scoring logic are explained; the items were widely field tested on a representative sample of test-takers; and further, empirical evidence is provided which demonstrates that VAT scores are structurally reliable indications of test-taker ability in MSA, suitable for high-stakes decision-making.

Finally, the concurrent validity evidence in this report shows that VAT scores are useful predictors of oral proficiency interviews as conducted and scored by trained human examiners. With a correlation of 0.87 between VAT scores and IRT-scaled ILR-OPI scores, the VAT accurately predicts more than 75% of the variance in U.S. government oral language proficiency ratings. Given the fact that the sample of ILR-OPIs analyzed here has a test-retest reliability of 0.91 (cf. Stansfield & Kenyon's (1992) reported ILR

<sup>1</sup> Correlation coefficients from test-retest data.

<sup>2</sup> The average of the correlations between all rater pairs

reliability of 0.90), the predictive power of VAT scores is about as high as the predictive power between two consecutive independent ILR-OPIs administered by active certified interviewers. Thus, the Versant Arabic Test efficiently predicts Interagency Language Roundtable Oral Proficiency Interview scores at close to the maximum possible level.

To assure the defensibility of employee selection procedures, employers in the U.S. follow the Equal Employment Opportunity Commission's (EEOC's) Uniform Guidelines for Employee Selection Procedures. These guidelines state that employee selection procedures must be reliable and valid. The above information provides evidence of the reliability, validity and legal defensibility of the Versant Arabic Test in conformance with the prescriptions of the EEOC's Uniform Guidelines.

**Authors:** Jared Bernstein and Masanori Suzuki.

**Principal developers:** Waheed Samy, Naima Bousofara Omar, Jian Cheng, Eli Andrews, Ulrike Pado, Mohamed Al-Saffar, Nazir Kikhia, Rula Kikhia, and Linda Istanbuli.

## 9. About the Company

**Pearson:** Ordinate Corporation, creator of the Versant tests, was combined with Pearson's Knowledge Technologies group in January, 2008. The Versant tests are the first to leverage a completely automated method for assessing spoken language.

**Ordinate Testing Technology:** The Ordinate automated testing system was developed to apply advanced speech recognition techniques and data collection via the telephone to the evaluation of language skills. The system includes automatic telephone reply procedures, dedicated speech recognizers, speech analyzers, databanks for digital storage of speech samples, and scoring report generators linked to the Internet. The Versant Arabic Test is the result of years of research in speech recognition, statistical modeling, linguistics, and testing theory. The Versant patented technologies are applied to Pearson's own language tests such as Versant English Test and Versant Spanish Test and also to customized tests. Sample projects include assessment of spoken English, assessment of spoken aviation English, children's reading assessment, adult literacy assessment, and collections and human rating of spoken language samples.

**Pearson's Policy:** Pearson is committed to the best practices in the development, use, and administration of language tests. Each Pearson employee strives to achieve the highest standards in test publishing and test practice. As applicable, Pearson follows the guidelines propounded in the Standards for Educational and Psychological Testing, and the Code of Professional Responsibilities in Educational Measurement. A copy of the Standards for Educational and Psychological Testing is available to every employee for reference.

**Research at Pearson:** In close cooperation with international experts, Pearson conducts ongoing research aimed at gathering substantial evidence for the validity, reliability, and practicality of its

current products and at investigating new applications for Ordinate technology. Research results are published in international journals and made available through the Versant website.

## 10. References

- Abdel-Massih, E. (1975). *A Sample Lexicon of Pan-Arabic*. Ann Arbor, Mich., Center for Near Eastern and North African Studies.
- Bachman, L. F. & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bull, M & Aylett, M. (1998). An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In R.H. Mannell & J. Robert-Ribes (Eds.), *Proceedings of the 5th International Conference on Spoken Language Processing*. Canberra, Australia: Australian Speech Science and Technology Association.
- Caplan, D. & Waters, G. (1999). Verbal working memory and sentence comprehension. *Behavioral and Brain Sciences*, 22, 77-126.
- Carroll, J. B. (1961). Fundamental considerations in testing for English language proficiency of foreign students. *Testing*. Washington, DC: Center for Applied Linguistics.
- Carroll, J. B. (1986). Second language. In R.F. Dillon & R.J. Sternberg (Eds.), *Cognition and Instructions*. Orlando FL: Academic Press.
- Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.
- Cutler, A. (2003). Lexical access. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science. Vol. 2, Epilepsy – Mental imagery, philosophical issues about*. London: Nature Publishing Group, 858-864.
- Jescheniak, J. D., Hahne, A. & Schriefers, H. J. (2003). Information flow in the mental lexicon during speech planning: evidence from event-related brain potentials. *Cognitive Brain Research*, 15(3), 261-276.
- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Levelt, W. J. M. (2001). Spoken word production: A theory of lexical access. *PNAS*, 98(23), 13464-13471.
- Linacre, J. M. (2003). Facets Rasch measurement computer program. Chicago: Winsteps.com.
- Linacre, J. M., Wright, B. D., & Lunz, M. E. (1990). A Facets model for judgmental scoring. Memo 61. MESA Psychometric Laboratory. University of Chicago. [www.rasch.org/memo61.htm](http://www.rasch.org/memo61.htm).

- Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, 40, 387-412.
- Luoma, S. (2003). *Assessing speaking*. Cambridge: Cambridge University Press.
- Miller, G. A. & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 2, 217-228.
- North, B. & Hughes, G. (2003, December). CEF Performance Samples: For Relating Language Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Retrieved September 26, 2008, from Council of Europe Web site: <http://www.coe.int/T/DG4/Portfolio/documents/videoperform.pdf>
- Ordinate (2003). *Ordinate SET-10 Can-Do Guide*. Menlo Park, CA: Author.
- Perry, J. (2001). *Reference and reflexivity*. Stanford, CA: CSLI Publications.
- Sieloff-Magnan, S. (1987). Rater reliability of the ACTFL Oral Proficiency Interview, *Canadian Modern Language Review*, 43, 525-537.
- Stansfield, C. W., & Kenyon, D. M. (1992). The development and validation of a simulated oral proficiency interview. *Modern Language Journal*, 76, 129-141.
- Van Turenhout, M. Hagoort, P., & Brown, C. M. (1998). Brain activity during speaking: From syntax to phonology in 40 milliseconds. *Science*, 280, 572-574.

## 11. Textbook References

- Peter F. Abboud, Ernest N. McCarus (Eds) (1983) *Elementary modern standard Arabic*. 2nd ed. New York: Cambridge University Press.
- Peter Abboud [and others] (1971) *Modern Standard Arabic: intermediate level*. Ann Arbor, Mich., Center for Near Eastern and North African Studies.
- Mahdi Alish (2000) *Ahlan wa Sahlan: functional modern standard Arabic for beginners*. New Haven, Yale University Press.
- Khalil Barhoum, (2006) *Spoken Arabic*. (Amel 120A/I) Stanford University Bookstore.
- Kristen Brustad, Mahmoud Al-Batal, Abbas Al-Tonsi. (2004) *al-Kitāb fī Taʿallum al-ʿArabiyya: A textbook for Arabic, Part One, Second Edition*. Washington, D.C.: Georgetown University Press.

Kristen Brustad, Mahmoud Al-Batal, Abbas Al-Tonsi. (1997) *al-Kitāb fī Taʿallum al-ʿArabiyya, al-juzʿ al-thānī: A textbook for Arabic, Part Two*. Washington, D.C.: Georgetown University Press.


Eckehard Schulz, Gunther Krah, Wolfgang Reuschel (2000) *Standard Arabic: An elementary-intermediate course*. New York: Cambridge University Press.

Defense Language Institute Foreign Language Center (DLIFLC)

- a. Arabic Basic Course: *Vocabulary List for From the Gulf to the Ocean*. (AD0493M) Book II, Lessons 1-12, November, 1998.
- b. Arabic Basic Course: *In-Context Vocabulary*, (AD0256S) Semester I, Units 1-7, July 2001.
- c. Arabic Basic Course: *Grammar Notes (AD 101)* (AD0280S) Semester I, Volume 2, Units 1-3, May 2005. Vocabulary pp. 83-102.
- d. Arabic Basic Course: *Grammar Notes (AD 102)* (AD0281S) Semester I, Volume 2, Units 4-7, May 2005. Vocabulary pp. 61-82.
- e. Arabic Basic Course: *Semester II Volume I (AD 201)* (AD0152S?) Units 8-12, Teacher Edition, June 2005. Vocabulary pp. 99-131.
- f. Arabic Basic Course: *Proficiency Enhancement Program, Reading Materials, Semester III (AD 301)* Student's Edition, [digital file] March 2006.
- g. Arabic Basic Course: *Proficiency Enhancement Program, Reading Materials, Semester III (AD 302)* Student's Edition, [digital file] March 2006.

## 12. Appendix: Test Materials




**Side 1 of the Test Paper:** Instructions and general introduction to test procedures. Note: These instructions are available in English and Arabic.



### Test Instructions

**Please read this before taking the test**

Versant tests are automated spoken language tests that are taken on the telephone or computer. If you would like to listen to a sample test, purchase a practice test, or view the test score after taking the test (if applicable), please visit [www.VersantTest.com](http://www.VersantTest.com)

Part	Instructions
<b>Before the Test</b>	<ul style="list-style-type: none"> <li>Carefully read this instruction page and the test paper. You may use a dictionary or ask someone for help if there are words or sentences that you don't understand.</li> <li>Choose a quiet location with a landline phone where you will not be interrupted during the test.</li> <li>Do not use a cordless phone, cellular phone, or VoIP phone (e.g., Skype™ or PC-to-phone services). Newer phones are generally better than older phones. Make sure that the phone is set to tone and not pulse.</li> </ul>
<b>Beginning the Test</b>	<ul style="list-style-type: none"> <li>To begin the test, call the phone number on the test paper using a landline push-button telephone.</li> <li>A recorded examiner's voice will guide you through each section of the test.</li> <li>Enter your Test Identification Number using the telephone keypad when the examiner's voice asks you to do so. This number is printed on the top right of your test paper.</li> <li>The examiner's voice will then ask you two questions: your name, and the city and the country you are calling from. If you are speaking too loudly or too quietly, the examiner's voice will tell you.</li> <li>The test begins when you say your name. If you hang up before you complete the test, the test cannot be graded. You cannot reuse the Test Identification Number.</li> </ul>
<b>During the Test</b>	<ul style="list-style-type: none"> <li>Hold the phone close to your mouth as shown in the picture below.</li> </ul> <div style="display: flex; justify-content: space-around; align-items: flex-end;"> <div style="text-align: center;">  <p>NO Too low, too far away</p> </div> <div style="text-align: center;">  <p>YES In front of mouth</p> </div> <div style="text-align: center;">  <p>YES A good distance</p> </div> </div> <ul style="list-style-type: none"> <li>Answer all questions smoothly and naturally in a clear, steady voice.</li> <li>If you don't know the proper way to respond to a test item, you can remain silent or say, "I don't know."</li> <li>Do not take notes or write during the test.</li> <li>When you hear, "Thank you for completing the test", you may hang up.</li> <li>If you wish, you may answer the optional questions at the end of the test. Your personal information will be kept anonymous.</li> </ul>

PEARSON

© 2011 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).



**Side 2 of the Test Paper:** Individualized test form (unique for each test taker) showing Test Identification Number, Part A; sentences to read, and examples for all sections.



## VERSANT ARABIC TEST

REMINDER: The test begins when you say your name. If you hang up before you complete the test, the test cannot be graded. You cannot reuse the Test Identification Number.

 **Call: 1.415.738.3800**

*Thank you for calling the Versant testing system.  
Please enter your Test Identification Number on the telephone keypad.  
Now, please say your name.  
Now, please say the city and country you are calling from.  
Now, please follow the instructions for Parts A through F.*

Test Identification Number

1234 5678

Expires: January 1, 2012

PART	TASK	TEST DETAILS
A	Reading	<p><i>Please read the sentences as you are instructed.</i></p> <p>1. مُحَمَّدٌ يُرِيدُ أَنْ يَتَعَلَّمَ السِّبَاحَةَ  2. سَيَزُورُ صَدِيقَهُ الَّذِي يَمْلِكُ بَيْتًا عَلَى شَاطِئِ الْبَحْرِ  3. ذَهَبَ إِلَى حَمَّامِ السِّبَاحَةِ فِي النَّادِي ثَلَاثَ مَرَّاتٍ فِي الْأُسْبُوعِ لِمُدَّةِ ثَلَاثَةِ أَشْهُرٍ  4. بَعْدَ أَنْ تَعَلَّمَ السِّبَاحَةَ شَارَكَ فِي مُسَابَقَةٍ وَفَازَ بِجَائِزَةٍ</p>
B	Repeat	<p><i>Please repeat each sentence that you hear.</i></p> <p>Example: a voice says, "الأعياد تساعد في التسلية والراحة"  and you say, "الأعياد تساعد في التسلية والراحة"</p>
C	Questions	<p><i>Now, please just give a simple answer to the questions.</i></p> <p>Example: a voice says, "الطالب يدرس ماذا يفعل الأستاذ؟"  and you say, "هو يدرس" or "يدرس"</p>
D	Sentence Builds	<p><i>Now, please rearrange the word groups into a sentence.</i></p> <p>Example: a voice says, "الأولاد ... الأكل ... يفضلون البيت"  and you say, "الأولاد يفضلون الأكل خارج البيت"</p>
E	Repeat	<p><i>Please repeat each sentence that you hear.</i></p> <p>Example: a voice says, "الأعياد تساعد في التسلية والراحة"  and you say, "الأعياد تساعد في التسلية والراحة"</p>
F	Passage Retelling	<p><i>You will hear several brief passages in Arabic. After each passage, you will hear a beep and then you will have 30 seconds to retell it in Arabic as best you can. Try to retell as much of the passage as you can in Arabic, including the important details.</i></p>

*Thank you for completing the test.*

PEARSON

Versant Arabic Test - Demo - 71 - 11111 - 6

© 2011 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s).

**Side 1 of the Score Report:** Summary of the test-taker's Overall score and subscores.

# SCORE REPORT


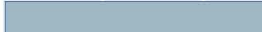
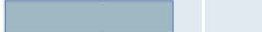

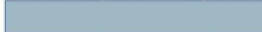
Versant Arabic Test



**Test Identification Number:** 12345678  
**Test Completion Date:** January 1, 2012  
**Test Completion Time:** 1:23 PM (UTC)

OVERALL SCORE

**48**

SKILL AREA	SCORE	20	30	40	50	60	70	80
<b>Overall Score</b>	<b>48</b>							
Sentence Mastery	53							
Vocabulary	37							
Fluency	44							
Pronunciation	55							

	DESCRIPTION
Overall	The Overall Score of the test represents the ability to understand contemporary spoken Modern Standard Arabic (MSA) and speak it intelligibly at a native conversational pace on everyday topics. Scores are based on a weighted combination of four diagnostic sub-scores. Scores are reported in the range from 20 to 80.
Candidate's Capabilities	Test-taker can follow a native-paced conversation and produce utterances using a variety of words and structures. Pronunciation is generally intelligible; test-taker can express some composite information on familiar topics to a cooperative listener.

© 2012 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s). Other names may be the trademarks of their respective owners.

For more information, visit us online at [www.VersantTest.com](http://www.VersantTest.com)

**PEARSON**





**Side 2 of the Score Report:** Detailed explanations of the test-taker's language capabilities.

# SCORE REPORT



**Test Identification Number:** 12345678

## EXPLANATION OF SUBSKILL SCORES

SKILL AREA	UNDERSTANDING THE SKILLS	CURRENT CAPABILITIES
Sentence Mastery	Sentence Mastery reflects the ability to understand, recall and produce Arabic phrases and clauses in complete sentences. Performance depends on accurate syntactic processing and appropriate usage of words, phrases and clauses in meaningful sentence structures.	Test-taker can understand, recall and produce many Arabic phrases and clauses in sentence context. Test-taker produces a range of meaningful sentences.
Vocabulary	Vocabulary reflects the ability to understand common words spoken in sentence context and to produce such words as needed. Performance depends on familiarity with the form and meaning of common words and their use in connected speech.	Test-taker has a limited understanding of basic spoken Arabic words, even when they are used in clear, simple speech.
Fluency	Fluency reflects the rhythm, phrasing and timing evident in constructing, reading and repeating sentences.	Test-taker speaks with adequate rhythm, phrasing and pausing. Hesitations and possible repetitions or omissions of words may result in an irregular speech rate and some disconnected phrases.
Pronunciation	Pronunciation reflects the ability to produce consonants, vowels and stress in a native-like manner in sentence context. Performance depends on knowledge of the phonological structure of common words.	Test-taker produces most vowels and consonants in a clear manner, although an occasional word may be unclear. Stress is placed correctly in most words. Speech is generally intelligible.

## About Us

Pearson creates unique technology for automated assessment of speech and text used in a variety of industry leading products and services. These include the Versant line of automated spoken language tests built on Ordinate technology, and WriteToLearn™ automated written summary and essay evaluations using the Knowledge Analysis Technologies™ (KAT) engine.

**To try a sample test or get more information,  
visit us online at:**

**[www.pearson.com/versant](http://www.pearson.com/versant)**

---

Version 1218D

© 2018 Pearson Education, Inc. or its affiliate(s). All rights reserved. Ordinate and Versant are trademarks, in the U.S. and/or other countries, of Pearson Education, Inc. or its affiliate(s). Other names may be the trademarks of their respective owners.