ICC

THE EUROPEAN LANGUAGE NETWORK

VERSANT

Spoken English Test

Quality Assurance Report

April 2008

International Certificate Conference, Berner Heerweg 183, 22159 Hamburg, Germany
www.icc-languages.eu
The International Certificate Conference (ICC) is an International Non-Governmental Organisation (INGO)
with participatory status at the Council of Europe

VERSANT Quality Assurance Report

Contents

Summary

Main Report

Certification

VERSANT Quality Assurance Report

Summary

1. Overview

Formerly known as the Spoken English Tests (SET), Versant tests demonstrate the highest standards of professional test development, founded on concepts of underlying linguistic competence popular in the United States. Versant offers a direct measure of speaking ability, where overall competence in spoken English is broken down into four diagnostic components – Sentence Mastery, Vocabulary, Fluency and Pronunciation. The Versant website gives a concise and accurate description for potential users or trainers:

> *Quick, Objective and Accurate Solution for Assessing Spoken Language Skills.*
> *Versant™ tests are the only completely automated tests of spoken languages. Using patented speech processing technology, Versant tests are delivered over a telephone or on a PC and scored by computer. Once administered, numeric scores and performance levels that describe the test-taker's ability to understand and speak the selected language are generated within minutes and can be viewed online. Scores are comprised of an overall score and diagnostic sub-scores.*

Versant refines a model of linguistic competence which can be said more or less to form the theoretical underpinning for scales such as that of the Foreign Service Institute/ Interagency Language Roundtable (FSI/ILR) and the related American Council for the Teaching of Foreign Languages (ACTFL) framework, and which has also been successfully implemented in English language tests, such as TOEFL and TOEIC. General reflections on Versant and its relationship to ACTFL/ILR and current developments in Europe as expressed in the CEF (*Common European Framework of Reference for Languages: Learning, Teaching , Assessment*, published by the Council of Europe) are included at 5 below.

The Versant test is at the forefront of currently accepted standards for test development, quality and research - indeed the quality of the supporting research is outstanding. The outlines and justification of test content, such as choice of vocabulary and structure, are clear and well-founded. The reliability of scores from the Versant programme meets or exceeds standards expected from human raters, and the explanation of how scores are generated and what they mean are clear for testing professionals and made accessible in the form of user-friendly score reports. In short, the Versant test is innovative, reliable and user-friendly.

2. Recommendation

The recommendation is that the Versant test be given a Certificate of Quality and Best Practice to acknowledge its excellent practice, particularly in the following areas:

- Describing the theoretical underpinning of the test and making explicit underlying assumptions about language ability and testing
- Outlining and justifying test content
- Explaining how scores are generated and what they mean
- Field testing and statistical evaluation of test results
- Reliability of scores
- Ability of scores to predict overall language competence
- A high correlation with scores given by human raters
- Wide and easy test availability
- Convenience of test administration
- Speed of result reporting
- Innovation in developing a fully viable machine-marked speaking test

The text of the ICC Certificate of Quality and Best Practice is at the end of this report.

Main Report

3. Nature of the test

The Versant test is a 10-minute test of spoken English delivered by telephone or computer and automatically scored. It aims to measure *the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English*. Including the time taken for the recorded prompts, the recorded samples play for between 14 and 15 minutes.

Test Sections

The test consists of 5 sections: Reading, Repetition, Short answer questions, Sentence builds, and Open questions. The first four sections include 58 items which are scored, and the last section includes three items which are not scored.

Reading
This section tests pronunciation and fluency.

Candidates read aloud sentences selected from those printed on the candidate prompt sheet. Candidates are able to study the prompt sheet and the sentences before the test begins. The sentences are arranged in groups of 4. Each group forms a connected paragraph, which is intended to assist in contextualizing the material, though the sentences for reading are selected in random order.

Repetition
This section measures listening ability, pronunciation, fluency, and sentence mastery.

Candidates hear sentences and repeat them. The underlying theory is that as language proficiency advances, so does the size of linguistic unit (i.e. number of words making up a phrase) that can be processed automatically. Longer and more complex sentences can therefore be processed and repeated by more advanced speakers, where speakers at a more elementary level are not able to recognise the individual words, group them into units and repeat them accurately within the time available.

Short answer questions
This section tests vocabulary.

Candidates respond with a word or short phrase. The questions ask for basic information or inferences based on time, sequence, number, word meaning or logic. They are intended not to assume any knowledge of Anglo-American culture, though occasionally this aim may not be entirely achieved (see notes below on recorded samples).

Sentence builds
This section tests pronunciation, fluency and sentence mastery.

Three words or short phrases are presented out of order. Candidates must re-arrange these to create a meaningful sentence. The section is a further test of candidates' ability to manipulate language in sense-groups. The larger a sense-group the candidate can manipulate, the higher their facility in spoken English. Our analysis shows that in practice the sentences do not invariably permit only one correct order – see notes on recorded samples at 6 below.

Open questions.
These questions are not currently marked by the computer and are not scored in the final calculation.

Three questions are presented with 20 seconds available for response to each. Candidates are required to state and briefly support an opinion rather than to analyse different sides of a complex issue at length and in depth. The topics selected appear to be relatively concrete and straightforward, focusing on common family and work issues. Questions are heard twice and the short response time available would tend to inhibit extended and complex replies.

Presentation of items

Sixty-one items are presented in each session, selected from a pool of some thousands of items. Items in the pool must have adequate discrimination, defined in terms of point biserial correlation with total score above a certain (unspecified) threshold. Items must also be understood and responded to appropriately by at least 90% of a reference sample of educated native speakers of English.

Presentation is not adaptive in that the selection of items is not based on or influenced by candidates' earlier responses. The selection of items is random, but stratified to ensure an equivalent distribution of item difficulties in each test version; an algorithm is used to ensure that items with similar content are not presented together - although this may perhaps be overridden by other factors such as appropriate item difficulty - see notes below on recorded samples at 6 below. Item difficulties are computed using a one-parameter Rasch (IRT) model.

Results from a study of 500 tests indicate that candidates repeating the test are unlikely to see more than one or two items again among the 58 items which are scored. The likelihood of exposure to repeated items will further reduce over time as the item pool increases.

4. Research and Statistical Information

An impressive body of research is described in the Versant Test Manual as well as in the Validation report for Phone Pass SET-10 (an earlier version of the Versant test) and in responses to queries made for this report. The comments below highlight a number of instances of best practice as well as raising one or two questions.

Test Construct

The underlying test construct is clearly spelled out. The test aims to measure facility in spoken English, described as

> *the ability to understand spoken English on everyday topics and to respond appropriately at a native-like conversational pace in intelligible English*

The ability to access lexis and encode speech into appropriate structures in real time (at a native-like conversational pace) depends on how much of the language a candidate has automated, to use Versant's word.

Versant deliberately focuses on the psycholinguistic aspects of spoken language performance rather than on the social, rhetorical and cognitive elements of communication. The test therefore emphasises the candidate's degree of automaticity with the language rather than his or her ability to recognise a context and produce a contextually appropriate cluster of phrases and vocabulary. The use of context-independent items is said to allow the most basic meanings of words, phrases and clauses (the building blocks of context-dependent meanings) to be tested. A high degree of context independence is further said to minimise the distorting effects of world knowledge and cognitive strategies on performance. Finally, with relatively context-independent items less time is spent developing a background cognitive/contextual schema for tasks. Within a necessarily limited amount of testing time, this allows candidates to spend more time demonstrating performance in speaking the language.

A view currently gaining favour in Europe is that language use is predominantly lexis-based and embedded in particular contexts rather than being based on an underlying competence which is then applied in different circumstances. While Versant aims to be context free, it is perhaps true that the no test item is quite as context free as the developers might like to think. An alternative explanation might be that the items are devised so that candidates can guess the context and respond appropriately with a minimum of scene setting. One of the underlying abilities being tested would then actually be the capacity to discern and switch contexts rapidly and repeatedly based on minimal clues. This type of mental agility would seem to have something in common with the abilities measured by intelligence tests. The question remains whether it is more natural/ authentic/ appropriate/ effective/ fair to test language embedded in a richer context.

Regardless of one's agreement or disagreement with the Versant developers' rationale for the testing methodology chosen, Versant for English is a meticulously developed test which is highly successful in ranking candidates' abilities.

Test Content

Versant for English shows evidence of very careful and thoughtful test design. The test measures both listening and speaking ability. The vocabulary is based on the 8000 most frequent word roots in the 1997 Switchboard Corpus, which is based on 3 million words taken from spontaneous telephone conversations, and the language structures in the test reflect those that are common in everyday English. Lexical and stylistic patterns used in test items are further based on conversations featuring 540 North Americans; a balanced representation of American English is given by geography and gender and reflects all major dialects. Although the test material is not directly based on English speakers from outside North America – and only North American voices appear to be used on the recordings - items are reviewed to ensure that they would be appropriate for test takers trained to standards other than US English. In particular, items are reviewed by UK and Australian linguists to ensure linguistic acceptability in those countries, and this measure is successful.

Test Scoring

As well as an overall score, Versant reports diagnostic subscores for Sentence Mastery, Vocabulary, Fluency and Pronunciation. Each response is analysed by a speech recogniser that is optimised for non-native speech, based on analysed responses from some hundreds of native speakers and some thousands of non-natives.

Content is scored according to the presence or absence of correct words in the correct sequence (Sentence Mastery and Vocabulary). Manner-of-speaking (Fluency and Pronunciation) scores are calculated by measuring the speed and phonetic accuracy of delivery. Score weighting is 50% each for content and manner-of-speaking.

Correlations between human and machine scores

The manner-of-speaking measures are scaled so as to optimally predict human judgments (i.e. whether a candidate speaks like a native or favourably-judged non-native). This scaling is partly based on 288 candidates (taking Phone Pass SET-10, an earlier edition of the Versant test) whose responses were graded both by human raters and by the computerised scoring system. Multiple human scorers were trained by master scorers and generated a total of over 26,000 scores (based presumably on approximately 288 candidates x 61 items per test = 17,568 responses). Human scores were compared against machine scores and scoring of the same items by multiple human raters allowed human inter-rater reliabilities to be calculated. As an example, human inter-rater reliabilities for fluency and pronunciation were between 0.74 and 0.86. Correlations between human and machine scores were in a similar range between 0.79 and 0.87.

These correlations indicate that agreement between machine scores and human scores is as good or slightly better than that between scores awarded by two different human graders. Even higher correlations between machine and human scores (as high as 0.97) are reported in the Versant for English Technical manual, but with a sample size of only n = 50, this report refers only to results from the earlier study, based on n = 288.

Scores are reported on a carefully devised logistic scale (non-native mean of 50, native score at 25th percentile of 75). Scores are generated between 10 and 90, but are reported with a minimum of 20 and a maximum 80, eliminating extreme scores considered to fall outside the effective measurement range of the test.

Test Validation

Prototype versions of Versant for English were administered in validation studies to over 4000 candidates (approximately 500 native and 3500 non-native speakers). From these a native norming sample of 376 literate adults and a non-native norming sample of 514 candidates (a stratified random sample with even representation for gender and native language) were selected. An extremely high correlation (of the order of 0.98) between results on these earlier editions and current editions of Versant indicate that inferences from these earlier validity studies remain valid for the current test.

A number of interesting features emerge from the validation studies and generally confirm test quality and good practice.

Spread of scores

There is a very clear distinction between scores for native and non-native populations. Most natives had very high scores (fewer than 5% scored below 68 on a scale of 20-80). Non-native learners showed a wide spread of scores across the range of the test (but only around 5% scored above 68). The test effectively separates native and non native speakers and is able to spread non-natives across an ability spectrum.

For understandable commercial reasons the test developers declined to provide data on all test takers to date. However, a random selection of 10,000 candidates from round the world confirmed the spread of non-native scores from the validation studies; as one might expect with such a large sample this showed a score distribution even more closely approximating the normal curve.

Correlations between subscores were significant but not perfect and follow a pattern that one might predict based on knowledge of the different components of speaking ability and their likely relationship. This would seem to support inferences made by the test developers about the different subscores representing different test constructs, at least in part. The subscores would at any rate appear to be sufficiently distinct to be useful diagnostically. The lowest correlation reported was 0.61, between fluency and vocabulary, which would instinctively seem to be different constructs. The highest was 0.92, between pronunciation and fluency, which would appear to be closely linked.

Correlations between Versant and other tests of speaking such as TOEFL Speaking are as high as 0.84 (n = 321). As expected, correlation with tests focusing on written academic English, such as TOEFL Reading are lower at 0.64 (n = 321).

Score Reporting

Scores are delivered in the form of a user-friendly score report. This indicates overall score on a kind of speedometer circle/pie chart and subscores as a bar graph.

The candidate's actual scores and the potential maximum and minimum scores for each criterion are shown in a very clear and intuitive way. A brief and clear explanation of what the test is meant to indicate and the score ranges is provided below the charts as well as an overview of what is meant to be measured by each subscore. The score report gives the band into which the candidate's score falls in each case and a brief clear descriptor of the level this is intended to represent.

In general, the score report is extremely clear with a surprising amount of information conveyed in a small space. Explanations are brief with technical terms reduced to an absolute minimum so as to make the descriptions accessible to an educated person who is not necessarily an expert on language.

Descriptors from the Common European Framework

The descriptors are drawn from the CEF and were related to the Versant scores in two main research studies.

The first study was carried out in 1999 and involved three independent raters and 120 students. Raters were carefully trained and standardised and scored candidate responses against CEF levels. Only the open response questions were rated by humans. Interrater reliability was 0.95 and correlation of the CEF levels assigned with Versant overall score 0.85. A preliminary score interpretation table was drawn up on the basis of this study. Because the preliminary study was based only on open response items, it was felt that it probably underestimated the ability of lower-level candidates, who might find these items inherently too difficult.

A second study in 2002 with four raters and 150 candidates used the other items in the Versant test to estimate fluency and pronunciation in addition to the information provided by the open questions and confirmed that test takers with scores under 50 had been underestimated. A new score transformation was calculated. The second study showed an interrater reliability of 0.94 and a correlation with Versant total scores of 0.94.

The studies used to interpret Versant scores in terms of CEF were carefully designed. However, some fundamental issues appear not yet to have been addressed.

- Is the sample generated by the Versant test adequate to assign CEF ratings?
- As well as not giving candidates at higher levels chance to demonstrate extended speech, the Versant test format might also not allow candidates at lower ability levels to demonstrate their abilities to the full because the test items do not consist of everyday tasks embedded in an authentic context of the type envisaged by the compilers of CEF.
- Were the scoring criteria used by the raters in the studies comparable with those that might have been used in normal face-to-face tests based on CEF, or were they unduly influenced by the specialised and restricted language sample available for rating?
- How significant is it that humans and machines appear to require different types of speech samples to assign ratings? (short, restricted range, predictable samples for computers; longer, wide-ranging open-ended samples for humans)
- Lower correlations with Versant scores than those reported above appear to have been obtained when the test was compared with extended face-to-face tests of speaking such as ILR-type interviews (correlation 0.75, n = 51). Might correlations between Versant and CEF ratings also be lower if the CEF ratings were also obtained from an extended face-to-face test?

Future studies may well shed light on these questions.

Correlation with other tests

Correlations with other tests tend to confirm that Versant has significant but not total predictive power with regard to scores on different tests. The results indicate a degree of commonality in what is tested but also significant differences in nature between other tests and Versant.

Correlations of 0.75 with TOEFL (n = 392) and 0.64 with TOEFL Reading (n = 321) suggest that perhaps Versant is not testing higher-level language of the type found on these tests.

Interestingly, the correlation with the tape-mediated TSE was 0.88 (n = 58). This perhaps reflects certain similarities in the testing philosophy underlying Versant and TSE.

5. General reflections on Versant and its relationship to CEF and ACTFL/ILR

The Versant test has been developed more within a framework which reflects US-based concepts than the mainstream models in Europe, especially those stemming from the Council of Europe. The CEF goes into considerable detail on the importance of identifying contexts in which language is used and in spelling out functions which the language user is able to carry out successfully. A further recent development in the language teaching and testing fields has been a growing focus on vocabulary and on breadth of lexical knowledge as a predictor of communicative success. Even such established American tests as the TOEFL have begun in their most recent versions to take on board some of these ideas. For instance, the TOEFL now incorporates tasks such as listening to a lecture and summarizing the main points orally or in writing. Such tasks enjoy face validity because they closely resemble activities which candidates would be called upon to engage in as students.

The chosen Versant testing methodology is meticulously developed, yielding a test which is highly successful in ranking candidates' abilities. The Versant test represents a considerable achievement in producing a fully viable machine-marked test of speaking.

Versant scores and CEF levels

Interpreting Versant scores in terms of CEF levels is certainly user friendly and useful. The Versant test can offer useful statistical predictions of test-takers' likely performance on CEF. However, the language elicited during the Versant test is almost entirely of a different type from that envisaged by the developers of CEF and appears to represent a fundamentally different view of language ability. It would seem that the Versant test does not generate speech samples that are entirely adequate to assign ratings to candidates based on CEF criteria. To a lesser degree, this observation applies also to ACTFL or LPI ratings. The significant philosophical differences between Versant and CEF will tend to limit the validity of interpretations of Versant scores that are based on CEF.

Although there are certainly limits to what Versant for English can measure and situations where a more context-specific test may be more appropriate, the items are devised so that candidates can guess the context and respond appropriately with a minimum of scene setting. One of the underlying abilities being tested would then actually be the capacity to discern and switch contexts rapidly and repeatedly based on minimal clues. This type of mental agility would seem to have something in common with the abilities measured by intelligence tests. The question remains whether it is more natural/ authentic/ appropriate/ effective/ fair to test language embedded in a richer context.

Diagnostic subscores

As well as an overall score, Versant reports diagnostic subscores for Sentence Mastery, Vocabulary, Fluency and Pronunciation, which would seem to be useful for teachers and learners. Statistically, the subscores appear to be measuring constructs which are sufficiently distinct to be useful diagnostically. Only overall scores were available for the recorded samples provided for analysis, so there are no comments on sample subscores.

Determining candidates' language proficiency

Versant for English appears to perform as well or better than human raters in estimating candidates' overall language ability, at least up to a certain ability threshold. The published Technical Manual claims that the test can make accurate predictions up to CEF C2 (perhaps ILR 2+/3 or ACTFL Advanced High/Superior). Further information supplied in response to queries for this report indicate that the upper limit of Versant's predictive power is ILR 3+ (ACTFL Superior+).

There is solid statistical evidence to support these assertions. The Free Response items are limited in scope and response time and are in any case not rated by the computer, effectively limiting candidates' scored responses to the level of words, phrases and isolated sentences. Given these limitations, the computerised scoring appears remarkably effective in predicting ability. However, as the Versant test does not call on candidates to produce extended discourse, some test users with strong views on face validity would perhaps be happier if the test made more modest predictive claims at higher levels. There is again the question whether candidates at lower levels might perform better with more contextual support than is available in the framework of Versant.

Computerised scoring and human CEF ratings

Three comparison studies between Versant machine scores and human ratings against CEF show correlations ranging from 0.84 to as high as 0.94. However, there might be some doubts about these results above CEF B2, since the recorded samples obtained from normal Versant for English tests would appear to be inadequate to establish a clear human CEF rating at these levels. When faced with a candidate to assess, examiners will naturally do their best to judge based on the information available, even when the speech sample is insufficient to make a fully justified decision based on the scoring criteria. The results of these studies should be interpreted with this in mind. CEF scores from more extended tests would likely still show a significant relationship with the machine scores, but probably at a lower level of correlation. This is indicated by comparisons against the wide-ranging ILR speaking interview (correlation of 0.75) and against tests such as TOEFL which focus on "higher-level" vocabulary and structures typically found in an academic setting (correlation 0.75).

Setting aside any reservations about the range of language tested or the naturalness of the testing context, it is certainly true that higher-level candidates' increasingly improved performance on simple (sentence-level) language tasks in terms of accuracy, fluency and pronunciation allows Versant's scoring algorithm to make useful predictions about their overall abilities. However, a correlation of 0.75 indicates a significant but by no means perfect relationship between two tests. Generally the power of one test to predict score variance on another test is estimated as the square of the correlation between the tests, i.e. with a correlation of 0.94, Versant scores would predict 88% of CEF score variation (where 0.94 squared = 0.88); with a correlation of 0.84, this reduces to 71% and with 0.75 correlation to only 56%. Useful as such predictions might be, tests specifically focused on higher level skills, English for Academic Purposes etc. would appear by no means to be displaced by Versant. Future studies will no doubt shed further light on Versant's effectiveness and predictive power at higher levels.

Proficiency levels and specialised contexts.

The test developers appear philosophically to favour the testing of general English ability, rather than more context-specific material. In addition, a test on the lines of Versant involves large-scale investment in development, item pre-testing and research, making it an unattractive proposition where candidate numbers are likely to be limited. Even the relatively short "free-response" questions appear currently to be too complex and open ended to allow computerised scoring. Increasing the amount of context provided to improve the subjective test experience for lower-level candidates would significantly increase the testing time needed to obtain a useful speech sample. These factors make it unlikely that similar tests will be devised in the near future to cover higher/lower levels or more specific content domains.

Bands for test scores

An informal investigation of recorded Versant test samples confirms that the bands into which overall scores are organised by the test developers appear to represent readily distinguishable levels of ability in spoken English. Frequently in recorded tests of spoken English sentence repetition and short responses are used to give examiners an initial estimate of candidate abilities which can then be refined by studying candidates' responses to more demanding questions, calling for longer stretches of free language production.

The Versant test appears to have mechanised this initial rough estimate process with remarkable success.

6. Analysis of Recorded Samples

Fourteen recorded samples were provided for analysis, presented in two sets. Each set consisted on a sample from each of the seven Versant score bands intended to represent CEF levels: Below A1; A1; A2; B1; B2; C1; and C2.

In both sets it was possible to see a clear progression of ability from the lowest-scoring sample to the highest. The samples were sufficiently distinct from each other to plausibly represent different levels of ability. Nevertheless, some reservations remain about Versant's interpretation of scores based on mapping against CEF levels, as referred to above.

Higher levels

Up to a level of CEF B1 (ILR 1+, ACTFL Intermediate Mid) the Versant speech samples provided were adequate to assign CEF/ACTFL levels to candidates with some confidence. Above this level, the absence of extended discourse exemplified in tasks such as narration, comparison and defending a position at length limited a rater's ability to effectively judge candidates' performance based on CEF or ACTFL criteria. It might be possible to get a sense of the quality of candidates' language, but not of its range.

Even making use of the open questions not used in the machine scoring, the available speech was a somewhat limited sample for human assessment purposes. Computer-scored responses are all at the word or sentence level. Even the open questions allow only a maximum of 20 seconds for response to each of the 3 questions (earlier test versions appeared to allow 30 seconds each on two questions which is still quite short).

In the Versant open questions, candidates are perhaps able to demonstrate CEF B1

> *Can link discrete simple elements into a connected sequence"*

and probably B2

> *Can produce stretches of language with a fairly even tempo. Clear, coherent, linked discourse, though there may be some jumpiness.*

It would not seem that candidates are required (or really able) to demonstrate C1

> *Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language…..Controlled use of connectors and cohesive devices.*

The Versant test gives even less scope for C2

> *Can express him/herself spontaneously at length with natural colloquial flow. Consistent grammatical and phonological control of a wide range of complex language, including appropriate use of connectors and other cohesive devices*

At higher levels CEF focuses on candidates' ability to produce extended discourse, while such discourse is largely absent from the Versant test. This would appear to limit the compatibility of Versant with CEF, and in particular Versant's ability to satisfactorily predict CEF scores at higher levels (above B2)

Lower levels

Another question is whether low-level candidates given more contextual support might have been able to provide richer speech samples than those elicited by Versant. Versant task types do not appear to be designed to elicit language at CEF A1

> *Can introduce him/herself and others and can ask and answer questions about personal details such as where he/she lives, people he/she knows and things he/she has*

or A2 – items referring to personal and family information, shopping, local geography, employment, matters of immediate need or B1 – situations likely to arise when travelling, experiences, events, dreams, hopes, ambitions, plans.

Basis of comparisons

Initial ability estimates were made in terms of the ACTFL guidelines. These perhaps lend themselves more to a yes/no decision making approach than do the CEF levels, which tend more to a continuum approach reflecting increasing variety and precision of language available to the candidate in performing common linguistic functions as well as expansion of the type of contexts in which these functions can be carried out. For example, ACTFL criteria are of the following kind when distinguishing ACTFL Novice/Intermediate or Intermediate Low/Intermediate Mid respectively.

> *Can understands/does not understand enough to carry on a face-to-face conversation*
> *Can carry on a simple face-to-face conversation inaccurately with very basic vocabulary/with an expanded vocabulary, but still inaccurately*

Candidates were also rated against (the analyst's) expectations of performance within the various CEF levels. Where information was lacking on performance against CEF criteria, there was occasional extrapolation based on (the analyst's) experience of the relationship between ACTFL and CEF.

The fact that ACTFL places considerable emphasis on the accuracy and fluency of language production and the extent to which language processing is automated is philosophically closer to the testing strategies of the Versant developers than to CEF. This perhaps makes it easier to estimate ACTFL/ILR levels as opposed to CEF levels on the basis of the Versant test samples. Nevertheless, similar objections could be raised that candidates are not asked to perform typical tasks such as

- ACTFL Intermediate (ILR 1): shopping or making travel plans
- ACTFL Advanced (ILR 2): narration, detailed description and comparison
- ACTFL Superior (ILR 3): analyzing different points of view in depth and defending a position at length.

Certainly it does not seem possible to estimate candidates' likely success at ACTFL Superior (ILR 3) tasks based on Versant. Similarly, the ILR 2+ Level (ACTFL Advanced High) is based on partly successful performance of Level 3 (Superior) tasks, so this level would also fall outside Versant's range.

However, even though narration is not specifically featured in the Versant test, the quality of language estimates probably provide enough information to justify assigning candidates to ACTFL Advanced (ILR 2).

Test presentation and content

Delivery

A variety of voices were used to present the prompts, but the same prompt appeared always to be presented using the same voice. Although the test material was verified for acceptability by British and Australian speakers of English, all the recorded voices were Americans.

Delivery, with a few exceptions, was at a fairly brisk conversational pace. One exception to this was one of the female speakers who spoke some of the sentence build items at an unnaturally slow pace and with a forced, almost aggressive delivery that seemed more geared to spelling the material out for the benefit of learners rather than to delivering it in natural style. The open questions were spoken twice and fairly slowly. It was not clear why this was so – it would seem to undermine the test's aim to measure the candidate's ability to understand and respond at a native conversational pace.

Content

In general the sentences and questions fulfilled the testers' aim of using a good range of everyday vocabulary and sentence structures. There was a fair amount of repetition of items from candidate to candidate – perhaps this is because the samples were drawn from the same question banks to facilitate comparison. Information provided in response to my enquiries would suggest that a candidate retaking the test would be unlikely to be exposed to more than a very minimal number of repeated items on the retake, however.

Test developers recognise that setting up naturalistic communication situations can be very time consuming. Nevertheless, a few of the items seemed a little unnatural, and reflecting an IQ test more than everyday conversation, for example:

> *If the Smiths' babysitter charges $5 an hour and they are out for 4 hours, how much would she charge?*
> *What type of expression is used when bumping into someone?*
> *To multiply quickly, would you use a typewriter or a calculator?*
> *The party was a disaster. Was it a success or a failure?*

Admittedly a competent speaker of the language should be able to answer any of these questions, and they are probably excellent discriminators, but they are open to some criticism on face validity grounds in that they call for candidates to talk about language rather than actually to communicate in a natural situation.

There were a few cases where items permitted more than one correct answer – for example the sentence build allows perhaps two equally plausible interpretations

> *some/food/buy* *Buy some food* or *Some buy food*

In Set 1 Candidate 5 both these seem equally acceptable

> *The weather this summer has been miserable*
> *The weather has been miserable this summer*

It was not specifically spelled out to candidates that the original order of words as given is not intended to be a possible correct answer: see the prompt for Candidate 1.6 which is already an acceptable English sentence in the order first given

> *Let's meet to play basketball at the park*

It is not clear in these cases if both options would be accepted by the scoring mechanism and whether the availability of a wider range of acceptable responses would reduce the usefulness of the item.

In general, the test material seemed fairly universal (assuming an audience of fairly educated adults from the type of socio-economic background likely to study English as a foreign language). A few questions seemed to require culture-specific knowledge, in spite of efforts to eliminate this – for example about picking guitar strings. Other topics such as winter or golf are perhaps not universal to all cultures. Dandelions do not grow in all parts of the world.

Occasionally the selection algorithm appeared to permit questions with similar content to appear together – for example in Set 1 No. 4, short answers, there were two beach questions together – about whale watching and building sandcastles.

It is interesting to note that even the best candidates in the samples do not ace the test with full marks, both scoring 79 out of a possible 80. There are still minor flaws and places where responses might be improved.

7. Comments on individual candidates by the analyst

The following brief notes on individual samples may serve to justify the observation above that the candidates on the Versant test show a clear progression in ability with increasing scores. They also illustrate the problems in assessing candidates at higher (and perhaps also at the lowest) ability levels. The differences between ACTFL/ILR, CEF and Versant seemed particularly apparent in the second set of samples.

Set 1.1 Below A1 (20-25) Score 23 Japanese Woman

Has some ability to read English and some knowledge of basic words. Speech emission is halting and pronunciation typically Japanese. Repetition is immediately a problem – unable to keep up with prompts and lapses into silence after a few items. Mostly just omits sentence builds. Has very little language internalised.

ACTFL Novice Mid range - maybe Novice Low (ILR 0); CEF Below A1.

Set 1.2 A1 (26-35) Score 30 Spanish Man

Pronunciation is largely incomprehensible with thick accent and delivery very stilted. Comprehension is quite limited. Re-interprets misheard words quite creatively in repetition. Mostly unable to respond on short answers. Quite a lot of miscomprehension. Creates some sentences in the builds section but substitutes many elements for those given. Keeps going but is relatively off topic on open questions. In general quite persistent, but inhibited by poor comprehension. Might well do much better with stretches of more familiar material.

ACTFL Novice High (ILR 0+); CEF A1.

Set 1.3 A2 (36-46) Score 38 Chinese Woman

Quite clear pronunciation and sounds as if she understands what she reads. Repetition is confident and rapid with good sentence rhythm but details are fudged. Seems to know some words but has little grammar internalised. Short answers show some comprehension, but responses are often unfocused. Words she misses seem to be more unusual ones rather than the most basic. Understands the sentence build task and can do the easier ones. Loses track of longer sentences. Fairly confident – doesn't have an outstanding vocabulary but probably understands enough to keep a conversation going.

ACTFL Intermediate Low (ILR 1); CEF A2 appropriate

Set 1.4 B1 (47-57) Score 54 French Woman

Smooth and fairly confident delivery with reasonable sentence tune but noticeably French pronunciation. Can repeat fairly well but fudges some small details such as final –s. Substitutes equally plausible but simpler alternatives such as "before we do the test" for "before taking the test". More complicated sentences impose some strain and she has surprising difficulty with "The plane leaves at 9.15 a.m." Comprehension and response is generally appropriate on short questions, though some of the more unusual ones cause problems. Takes a while getting into the sentence builds – perhaps isn't clear about the task. The longer ones cause strain. Responds fairly well to open questions. Demonstrates a decent vocabulary range, but not much of the grammar seems internalised, which lowers the grade.

ACTFL Intermediate Mid rather than Intermediate High (ILR 1 to possibly 1+); CEF B1.

Set 1.5 B2 (58-68) Score 64 French Man

Shows typical French pronunciation with a fair amount of words mis-stressed and some odd sounding vowels. He seems to have little problem understanding and repeating the sentences. There are a couple of slips and some loss of fine detail. His poor pronunciation probably makes him sound worse than he really is. Comprehension on short answers is generally accurate with only a few minor problems. Some of the less common vocabulary is unfamiliar to him – peninsula/bay; cherry stone. Sentence builds are well done with a fairly sound grasp of the grammar. His accuracy goes down on the last 3 or 4 more difficult items. Answers on open questions are focused though not exceptionally long. Probably demonstrates enough internalised language for ACTFL Intermediate High. (ILR 1+). Might possibly even be ACTFL Advanced – not enough evidence of extended speech, narration, comparison etc. to be confident of this. It is also true that candidates' speech tends to decline in focus and accuracy in longer stretches of production. The speech that was available did not seem to reflect "pausing for grammatical and lexical planning and repair may be very evident", so candidate is possibly CEF B2 rather than B1, though it is somewhat difficult to tell without more extended speech ("stretches of language"). Language range (vocabulary and grammar) is certainly better than 1.4 candidate. It is perhaps still marginally possible to judge on ACTFL/CEF criteria based on the available speech sample.

ACTFL Intermediate High. (ILR 1+); possibly ACTFL Advanced; possibly CEF B2 rather than B1.

Set 1.6 C1 (69-78) Score 74 Dutch Man

Fluent and confident reading. Pronunciation is natural if not entirely native – perhaps slightly jerky. Repetition is good but not perfect with a few alterations and substitutions (prepositions + pronouns, "your voicemail" for "his voicemail" "stayed the same" for "remained the same" etc. Pronunciation was slightly slurred and intonation not totally native but overall easy on the ear. Short answers were well done with only 3 or 4 of the more difficult items missed. Sentence builds were fairly capably handled – candidate was in control of the language to the extent that he realised "Let's meet/ to play basketball/ at the park", although correct in itself should by the assumed conventions of the item format be changed to "Let's meet at the park to play basketball." The candidate gave a clear and accurate response to the open questions and used up the available time. He would probably have had rather more to say given more time. Speech here was generally correct though vocabulary was restricted to fairly common items and responses were not particularly focused. Based on the accuracy of speech and use of fairly basic vocabulary here, candidate would appear to be ACTFL Advanced. (LPI 2). Evidence was not sufficient to support a rating of Advanced High (2+), since the candidate was not challenged to produce more sophisticated vocabulary and more complex sentence structure in extended responses. CEF was more difficult to rate. Candidate certainly did not show the pausing, reformulation and repair characteristic of B1 and would appear to be at least B2. Not enough language was available to determine clearly if C1 would be warranted.

ACTFL Advanced. (LPI 2); at least B2.

Set 1.7 C2 (79-80) Score 79 German Woman

Reading was clear and accurate. Pronunciation does not sound totally native, but is otherwise pretty well flawless. Repetition was close to perfect, with only a few minor simplifications to suggest that the speaker does not habitually use the full range of English structures – "we had" for "we've had" ; "we have to" for "we'll have to". The short questions seemed pretty easy for the candidate, though 3 or 4 questions featuring more unusual vocabulary were missed. Sentence builds generally posed no problems, though again there were one or two indicative places where the speaker substituted her own version of the language. "The girl puts on a red dress" for "The girl put on a red dress" and "The cost will increase when fewer than 2 will sign up" (for "when fewer than 2 sign up". Responses to the open questions were fluent and precise and reflected very much the language one might expect from a solid speaker at ACTFL Advanced (perhaps Advanced Mid) (LPI 2). I would have little hesitation assigning an Advanced rating, though I would not feel able on the basis of this sample to assign Advanced High (2+) or better. The open questions do not appear to call for language at ACTFL Superior/LPI 3 and therefore do not allow one to judge whether the candidate is capable of going beyond Advanced/LPI 2. This candidate appeared more confident and at ease with the language than 1.6. On CEF, the candidate would seem to be at least B2. My impression is that the level is probably C1 (fleeting evidence that the candidate's knowledge of the everyday structures of English is not entirely complete might hint that C2 could be too high). It should be stressed that this is based purely on subjective "quality of delivery" impressions rather than on demonstrated performance against CEF criteria.

ACTFL Advanced (perhaps Advanced Mid) (LPI 2); at least B2, probably C1.

Set 2.1 Below A1 (20-25) Score 23 Turkish Woman

Pronunciation is very noticeably foreign. Reading is fairly accurate but very laboured. It takes the candidate a long time to get into the repetition section with no response at all for most items until the last couple, where she is able to repeat only a few isolated words. Most of the short response questions are answered incorrectly. The candidate gets about 4 items correct. Sentence builds are clearly too difficult for her. She manages only to repeat isolated words from the prompts. Open questions again are too difficult. For each question, the candidate begins "I like" and offers only one or two words beyond this. In ACTFL terms, the candidate is clearly a Novice (ILR 0). The test sample suggests Novice Low, though tasks more appropriate for the level might conceivably enable her to demonstrate Novice Mid. In CEF terms, the candidate is probably below A1, though again might demonstrate A1 given more suitable tasks.

ACTFL Novice (ILR 0); below CEF A1.

Set 2.2 A1 (26-35) Score 34 Telugu Man
Candidate has a heavy accent but seems to be able to read English aloud. Repetition reveals some serious comprehension problems with most items missed or significantly distorted. About a dozen short answer questions are missed completely or answered incorrectly. Sentence builds are mostly answered incorrectly. On the open questions, the candidate tends to recount his own experiences on topics vaguely related to one or more words in the prompt rather than actually responding to the question asked. Knows some English, but generally weak comprehension.

ACTFL Novice High (ILR 0+); CEF A1.

Set 2.3 A2 (36-46) Score 40 Filipino Woman
Reads and speaks confidently, but pronunciation is quite hard to understand. Misses or significantly distorts most of the repetitions. Candidate also misses about a dozen short answers. On the sentence builds, she gets the order of elements right but tends to transform everything into her idiosyncratic version of English grammar. Responses on the open questions are fairly unfocused. In general the candidate seems to know basic words and some of the more unusual ones and have a fair amount of confidence and persistence in speaking which fits a profile of ACTFL Intermediate Mid (ILR 1). CEF level is hard to judge because generally the test questions do not call for the candidate to produce "main repertoire associated with more predictable situations" or produce much of a sequence of speech. Candidate seems certainly A2, might possibly be B1.

ACTFL Intermediate Mid (ILR 1); CEF A2, possibly B1.

Set 2.4 B1 (47-57) Score 53 Persian Woman
Candidate reads well but somewhat hesitantly. Repetition poses difficulties as the sentences get longer and there seem to be a few vocabulary/comprehension problems. This pattern is continued in the short answers with about a dozen items missed to a greater or lesser degree. Sentence builds are reasonably done but with some fudging ("parents" for "uncle" "Charlie is thinking to learning of playing of the guitar" etc. The open response items are answered concisely and accurately but with vocabulary remaining at a fairly basic level. This candidate is interesting in that what she does know, she seems to know and use quite solidly and accurately, but there appear to be significant gaps in her knowledge. CEF is harder to judge than ACTFL.

ACTFL Intermediate High (ILR 1+); perhaps a strong CEF B1.

Set 2.5 B2 (58-68) Score 64 Marathi Man
Reads clearly and naturally with a noticeably Indian-style accent. Repetition reveals a number of omissions and changes ("the Richard story" for "Richard's stories" etc) and the candidate begins to get a bit tangled up as the sentences get longer. He misses around 8 short answers and tends to produce one word responses with no articles. Sentence builds start well, but the longer items start to pose problems. Open questions generally produce good responses, though there is some evidence of strain (for instance in frequent repetition of the word "help"). Generally speech seems more typical of ACTFL Intermediate High (ILR 1+) than Advanced (ILR 2). ACTFL and CEF are by no means exactly equivalent. Candidate seems a reasonable match for CEF B1.

ACTFL Intermediate High (ILR 1+); CEF B1.

Set 2.6 C1 (69-78) Score 71 Tamil Man

Pronunciation sounds virtually native US, though slight jerkiness and over-articulation occasionally detract from the overall effect. Repetition is generally very accurate though in the last half dozen items a few things begin to be missed. Responses to short answer questions are generally good, missing around 4 items. The candidate tends to respond with 2 or 3 words even where this is not totally idiomatic and he seems to avoid using pronouns (for instance "above the house" where one might expect a native speaker to say "above it"). Sentence builds are generally well handled. Responses on open questions are interesting. The candidate shows some weaknesses – for example he repeats "for the simple reason that" in every answer, but there is evidence of structures and vocabulary reflective of ACTFL Superior (ILR 3), such as

> *It is rather more important for the student to be dynamic rather than specialised*
> *The kind of life you can offer your children*
> *Rather than in a suburban area where you might get a peaceful life*
> *It would affect your relationship if you are not in a position to pay the money back at the promised time*

Evidence is incomplete, but there are hints here of a possible ACTFL Advanced High (ILR 2+). Evidence of effective use of connectors and cohesive devices (within the limited scope of the open questions) might suggest CEF C1, even in the absence of really extended speech.

ACTFL Superior (ILR 3); possible ACTFL Advanced High (ILR 2+); CEF C1.

Set 2.7 C2 (79-80) Score 79 Hindi Woman

Pronunciation is excellent if occasionally slightly bumpy. Repetition is near perfect until the last few items where there are one or two substitutions, such as "protected" for "protective". Short answers are answered confidently and appropriately with answers tending to be a single word. The candidate does not use articles as frequently as might be expected from a native speaker, otherwise speech is close to perfect here. Sentence builds are well done though one or two are missed. Open responses are accurate but lack more complex vocabulary characteristic of ACTFL Superior (ILR 3). There are one or two mispronunciations ("comfortable"). Based on the sample the candidate seems clearly ACTFL Advanced (ILR 2), but there is no evidence of speech at ACTFL Advanced High/Superior (ILR 2+/3). CEF is difficult to judge. The candidate seems clearly at least B2. Evidence for higher levels is lacking.

Clearly ACTFL Advanced (ILR 2); at least CEF B2.

Certification

This report has been submitted through the Quality Assurance Scheme to the Executive Board of the ICC and accepted.

The Executive Board notes the many points of excellence mentioned in the report and the suggested areas for further research and development. The recommendation of the Executive Board is that a Certificate of Quality and Best Practice be issued with the text shown below. A facsimile of the Certificate is attached as a draft for discussion before final release.

ICC

Certificate of Quality and Best Practice

has been awarded to

Pearson Education

and

Ordinate Technology

for

VERSANT

Spoken English Test

This Certificate of Quality and Best Practice has been awarded to Pearson Education and Ordinate Technology for he development of he Versant English language tests adminuste3red by computer or telephone which after appraisal by experts appointed by the ICC – Quality Assurance Scheme are recognised as exemplars of Quality and Best Practice, particularly with regard to

- Test development and research, including field testing and statistical evaluation
- Innovation in developing a fully viable machine-marked speaking test
- Reliability of scores and high correlation between machine and human raters
- Ability of scores to predict overall language competence
- Convenience of test administration
- Accessibility and transparency for test takers

This certificate is valid for a period of five years from the date of issue, provided no major changes are made in the existing scheme.

(Signed)

Director **President**

Serial No. 1008

April 2008