

Scoring Criteria, Methods and Reliability

Introduction

In this chapter we present a detailed description of the scoring criteria, methods and reliability of the CAEL Assessment. A team composed of test developers and experienced ESL/EAP teachers at the Language Assessment and Testing Research Unit (LATRU) in the School of Linguistics and Applied Language Studies at Carleton University developed the score criteria for each sub-test of the CAEL Assessment. A concise but comprehensive description of the development of the CAEL Assessment can be found in Fox, Pychyl, & Zumbo (1993). The score criteria and methods are continually monitored and occasionally revised in order to maximize both efficiency and reliability. The score criteria are provided for each sub-test as well as for the Overall Result as a set of band scores (10 - 90) which range from *Very Limited Proficiency* to *Expert Proficiency*. A detailed description of the standard procedures used in scoring every CAEL Assessment is also presented. Since the scoring procedure is a little different for each sub-test, the score criteria and methods are presented separately for each sub-test of the CAEL Assessment (Writing, Listening, Reading, and Speaking) and then for the Overall Result.

In the *Standards for Educational and Psychological Testing*, reliability is defined as "The degree to which test scores are consistent, dependable, or repeatable, that is, the degree to which they are free of errors of measurement" (AERA, APA, & NCME, 1985). CAEL test developers endeavour to minimize all possible sources of error in the administration of the test and in the scoring and reporting of test results. Test developers recognize that the unique features of the CAEL Assessment necessitate the gathering of some very specific types of reliability evidence. Given that the CAEL Assessment is comprised principally of constructed-response items and that the score interpretations are criterion-referenced, the consistency of the scoring procedures is one important aspect of the reliability of this test. Constructed-response items require a greater degree of subjective rating than do selected response item types such as multiple choice items. For this reason, inter-rater reliabilities are routinely calculated for each of the CAEL sub-tests. Measures of the internal consistency of the test items (i.e., split-half reliabilities), often cited in test manuals, are not really meaningful given the test specifications, item types, and interpretations made from CAEL Assessment results (for a detailed discussion of this issue see Zumbo, Fox & Pychyl, 1993). Accordingly, this chapter presents the inter-rater reliability evidence that is collected on an on-going basis for each sub-test of the CAEL Assessment.

The second unique feature of the CAEL Assessment that necessitates the collection of reliability evidence is the test is a topic-based integrated assessment available in a number of different versions. In accordance with the principles of fair and ethical testing as defined by the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985), the CAEL test development team is committed to ensuring the comparability of these different test versions.

This chapter presents evidence from a number of sources, which demonstrates the consistency of scores across different test versions.

Writing Performance - Score Criteria, Methods and Reliability

The writing band scores that have been developed for the CAEL Assessment are the result of a collaborative effort between CAEL test developers and experienced ESL/EAP teachers. The theoretical background knowledge and depth of experience of these ESL/EAP teachers is an invaluable resource for the test development team. These teachers are well versed in the curriculum and teaching methods for courses ranging from intensive English language courses for students with little or no prior English language training to the final stages of advanced academic English language courses. Test developers at LATRU work closely with the ESL/EAP teachers to develop pilot testing procedures that include students at various levels of English language proficiency. Pilot testing includes students from the full range of ESL/EAP course levels, as well as students admitted to first-year university courses who have already satisfied Carleton University's English language requirements. Some pilot testing is also conducted with students at the senior high school level and at other universities and colleges within Canada and abroad. This extensive pilot testing helps test developers to ensure that the full range of academic writing proficiency described in the score criteria is adequately sampled in the test development process. The collaboration between the ESL/EAP teachers and the test development team has resulted in well developed and articulated criteria for the Writing Performance sub-test of the CAEL Assessment.

The score criteria for Writing Performance is presented in Table 4.1. The band scores encompass a wide range of writing proficiency - from the *Very Limited Writer* (10-20) to the *Expert Writer* (80-90). These band score descriptions are included with the CAEL Assessment Score Report which is given to every CAEL test taker. This report provides the test taker with a description of their own performance as well as a description of the full range of band scores. This information is a valuable resource for CAEL test takers and the institutions which receive CAEL Assessment Score Reports.

Table 4.1: Writing Performance Band Score Criteria

Band Score	Score Criteria
10-20	<p>Very Limited Writer</p> <ul style="list-style-type: none"> • is generally unable to express ideas effectively in writing • uses very restricted and/or ungrammatical language • uses words randomly and without overall coherence
30	<p>Limited Writer</p> <ul style="list-style-type: none"> • attempts to write something which is related to the topic but the writing is not predictable • uses restricted and/or ungrammatical language • seems to understand the topic, but is unable to develop ideas because language constrains or distorts expression
40	<p>Marginally Competent Writer</p> <ul style="list-style-type: none"> • makes links among ideas and addresses the topic but the writing lacks clarity and cohesiveness • displays elements of control in the writing (e.g. a thesis statement, an introduction and conclusion) but internal coherence is lacking • uses little or no support (i.e. quotations, examples, etc.) to develop the thesis
50	<p>Competent but Limited Writer</p> <ul style="list-style-type: none"> • addresses the topic to a degree but with somewhat limited clarity and cohesiveness • uses some support to develop the thesis • control of the argument is limited by poor comprehension of the readings and lecture, and/or poor understanding of the requirements of academic writing
60	<p>Competent Writer</p> <ul style="list-style-type: none"> • develops a thesis using a range of support • uses language that is generally accurate but is constrained by a somewhat limited vocabulary • demonstrates general understanding of the requirements of academic writing
70	<p>Adept Writer</p> <ul style="list-style-type: none"> • responds readily to the demands of the topic and presents information clearly and logically • uses the readings and lecture effectively to support the thesis • demonstrates understanding of the requirements of academic writing
80-90	<p>Expert Writer</p> <ul style="list-style-type: none"> • demonstrates mastery of appropriate, concise, and persuasive academic writing • writes with authority and style

The scoring method used for the Writing Performance sub-test of the CAEL Assessment was developed by test developers within LATRU (see Fox & Soucy 1991) and is referred to as the "collaborative read-aloud marking protocol." Earlier experience with traditional methods of the holistic assessment of writing (Diederich, 1967) highlighted the impact of marker fatigue on the consistency of ratings (Fox & Soucy, 1991). CAEL test developers sought to devise a rating system that would minimize factors such as marker fatigue in the rating of the Writing Performance sub-test. Research investigating many factors that influence the consistency of ratings for writing was carefully reviewed and used as a guide in the development of the collaborative read-aloud marking protocol. The resulting protocol is a unique procedure that contributes to both the reliability and validity of the CAEL Assessment. For this reason, the procedure described in some detail is included in this chapter.

The writing samples produced by CAEL test takers in response to the essay prompt are typically essays of one to three pages in length. CAEL Assessment essays are marked 'blind'. That is, essays are identified using only the test taker's identification number and initials. Raters are not aware of the specific characteristics (gender, first language etc.) of the test taker. All CAEL Assessment Writing Performance raters are experienced ESL/EAP teachers who have received specific training in the marking of CAEL Assessment essays. Version-specific marking guide descriptors and sample essays are used in the training of new raters and are also available to the raters throughout the rating process.

A team of three trained raters meets to begin the marking protocol. The first essay is selected and one rater is chosen to read the essay aloud. Raters are trained to read the essays in as smooth and fluent a manner as possible without revealing in any way (by changing inflection in the voice, editorial comments, etc.) what the reader thinks of the piece being read. The writing samples are read aloud in order to minimize the impact of surface features in the text such as spelling and handwriting on the score assigned. The impact of these surface features has been researched extensively and can be substantial. Raters are encouraged to place greater emphasis on other more substantial features (e.g., meaningfulness, cohesiveness, flow, etc.) of the essays and to discount these surface features when assigning a score. After listening to the essay, each rater records brief descriptive comments indicating their impression of the content, language and organization of the writing sample. Each individual rater also records a score for the writing sample. At this time, individual raters may ask for the writing sample to be read again or they may ask to see the piece of writing. After each individual rater has recorded their mark, the marks are revealed. If there is consensus on the marks, then the raters go on to the next paper. If there is a difference in the scores assigned, a discussion begins. In this case raters describe their rationales for the marks they have given and a consensus is reached. When the consensus has been reached, the final writing band score is recorded for the test taker. This collaborative marking process results in a consistent application of the score criteria that have been developed for the CAEL Assessment.

Rarely an essay is encountered for which the original marking team cannot reach a consensus. When this occurs a second team of raters is asked to review the writing sample.

The inter-rater reliability of this process is constantly monitored. A number of previously scored writing samples are re-marked at every rating session. The correlation between the first and second marking occasions is calculated. Because the scale used in marking the writing samples is an ordinal scale, a Spearman rank-order correlation co-efficient (ρ) is reported. The most recent correlation co-efficient is .962 ($n=178$). Because the inter-rater reliability data is collected on an ongoing basis, a number of teams of raters and all current test versions are included in this sample.

In addition to Spearman's ρ , the distribution of score changes for these 178 writing samples is shown in Table 4.2. The frequency of scores changed is shown as a percent. No change indicates that the band score assigned by the second team of raters was exactly the same as the band score assigned by the original team of raters. A change of one indicates that the band score assigned by the second team of raters was either one band score above or one band score below the band score assigned by the original team of raters. A similar description applies for the changes of 2 and 3 band scores. Table 4.2 indicates that 87.1 % of the writing scores are exactly the same across the two groups of markers, or they are within one band score of each other. Given the number of factors influencing the rating of essays, as reported in a number of research studies, CAEL test developers are confident that the reliability of the collaborative read-aloud marking protocol meets or exceeds the reliability of essay rating processes used in other contexts.

Table 4.2: Writing Performance Band Score Changes (n=178)

Band Score Difference	Freq. (%)
No change	87.1
Change of 1 band score	10.1
Change of 2 band scores	2.2
Change of 3 band scores	0.6

Listening Performance - Score Criteria, Methods and Reliability

The score criteria for the Listening sub-test are provided in Table 4.3. These criteria have been developed collaboratively by test developers and ESL/EAP teachers at LATRU in much the same manner as they were for writing performance, see description above. The principal difference between establishing the Listening Performance criteria and the criteria for Writing Performance lies in the fact that an initial raw score is calculated when the Listening Performance sub-test is marked. Raw scores are not generated for the Writing Performance. The raw scores for each version of the CAEL Assessment are then converted into standard band scores on the basis of the results from the pilot testing process. As is the case for the Writing sub-test, extensive pilot testing of students across a broad range of English language proficiency is conducted.

This procedure enables test developers to make accurate and consistent conversions of the raw scores generated from each version of the test into standard band scores.

The band scores shown in Table 4.3 reflect a wide range of listening proficiency from the *Very Limited Listener* (10-20) to the *Expert Listener* (80-90). As with the Writing Performance sub-test, these band score descriptions are included in the CAEL Assessment Score Report which is given to CAEL Assessment test takers. This report provides the test taker with a description of their own Listening Performance as well as a description of the full range of band scores.

Table 4.3: Listening Performance Band Score Criteria

Band Score	Score Criteria
10-20	Very Limited Listener Demonstrates very limited comprehension of lectures takes some meaning from individual words overall understanding is sketchy and random
30	Limited Listener Demonstrates limited and inconsistent comprehension of lectures makes sense of some sections of lectures by guessing overall understanding is limited
40	Marginally Competent Listener Demonstrates uneven comprehension of lectures is able to identify the meaning of some unfamiliar terms overall understanding is restricted
50	Competent but Limited Listener Demonstrates somewhat limited comprehension of lectures is able to process most lecture sections for general ideas, but misses or misinterprets specific details from time to time overall understanding is still somewhat restricted
60	Competent Listener is able to understand information regarding both main ideas and supporting details in lectures may lack some flexibility and miss some information compensates at times for missed information by drawing on overall understanding of what is being said
70	Adept Listener is able to understand lectures with apparent ease compensates strategically for unfamiliar vocabulary or terminology overall understanding is flexible and consistent
80-90	Expert Listener Demonstrates comprehension of lectures which is equal to that of experienced academic listeners Understands both main ideas and supporting details with ease is fully engaged by and interacts with the information being presented

Scoring of the listening component is conducted by trained raters using detailed marking keys. The marking keys provide explicit examples of acceptable responses and indicate the number of points to be allotted for each response. The marking keys are occasionally revised to reflect the range of responses that are provided by CAEL test takers. Partial scoring including half marks is allowed for all of the constructed-response items. A raw listening score is determined by summing the points for each of the items on this test component. The raw listening score obtained on each version of the CAEL Assessment is converted into a band score so that CAEL Assessment scores can be compared across versions.

Scoring of the Listening Performance sub-test of the CAEL Assessment is more objective than the scoring of the Writing Performance. However, a certain amount of rater subjectivity is always involved in the rating of constructed-response test items. For this reason, the inter-rater reliability of the Listening Performance is constantly monitored on a version-by-version basis. A number of listening tests are re-marked by individual raters at regular intervals. Because the raw scores are converted to a band score which is considered to be an ordinal scale of measurement, a Spearman rank-order correlation coefficient (ρ) correlation coefficient is calculated as a measure of the extent to which the two groups of raters provide similar scores. The most recent investigation of the inter-rater reliability of the Listening Performance sub-test resulted in a Spearman rank-order correlation coefficient (ρ) of .973 (n=178). With regard to the distribution of band score changes, 93.8% of the band scores remained the same. Given the consistently high correlations that result from this process, CAEL test developers are confident that the scoring procedures for the Listening Performance sub-test are reliable.

Reading Performance – Score Criteria, Methods and Reliability

The score criteria for Reading Performance are provided in table 4.4. The development of these criteria is very similar to the process described above for the Listening Component. As with the Listening Component, raw scores for Reading Performance have been converted into standard band scores on the basis of the results of extensive pilot testing involving students with a wide range of English language proficiency. The score criteria range from the *Very Limited Reader* (10-20) to the *Expert Reader* (80-90).

Once again, these band score descriptions are included in the CAEL Assessment Score Report which is given to every CAEL test taker. This report provides the test taker with a description of their own Reading Performance as well as a description of the full range of band scores.

As described earlier, each CAEL Assessment typically includes two readings, both of which are related to the topic of the test. Responses to the items for each reading are scored using detailed scoring keys in much the same manner as the scoring for the Listening Performance sub-test. The raw scores for the two readings are added together and the total raw score for Reading Performance is then converted to a band score.

Table 4.4: Reading Performance Band Score Criteria

Band Score	Score Criteria
10-20	Very Limited Reader is unable to read effectively takes some meaning from pictures, titles, random words, etc. may understand the main idea at times but misses almost all of the supporting details
30	Limited Reader reads with limited accuracy and fluency reads with some understanding of the main ideas but is unable to identify specific, relevant details is often unable to identify the meaning of unfamiliar terms from context
40	Marginally Competent Reader is unable to understand main ideas is restricted by limited vocabulary and a lack of familiarity with textural conventions reads more slowly than most academic readers
50	Competent but Limited Reader reads with understanding of the main ideas and is able to identify some relevant details reads more slowly and with greater effort than most academic readers may misinterpret information at times
60	Competent Reader Understands main ideas and is able to identify most relevant details reads more slowly and with greater effort than some academic readers is able to interpret information with some flexibility
70	Adept Reader reads academic texts with ease provided sufficient time is available demonstrates comprehension of academic texts which approaches that of experienced academic readers interprets information with flexibility
80-90	Expert Reader reads academic texts with ease demonstrates comprehension of academic texts which is equal to that of experienced academic readers understands both main ideas and supporting details with ease

The inter-rater reliability of the scoring for this section of the test is also monitored by re-marking a certain number of tests. After the raw score is converted into the band score, a Spearman rank order correlation co-efficient is calculated on the band scores from the two raters. The most recent investigation of the inter-rater reliability of the Reading Component resulted in a Spearman rank order correlation co-efficient of .949 (n=178). Frequency distribution of the number of band changes show that 91.6% of the remarks resulted in the same band score.

Based on this information, CAEL test developers are confident that the scoring methods used for the Reading Performance produce reliable results.

Speaking Performance - Score Criteria, Methods and Reliability

Test developers at LATRU and ESL/EAP teachers at the School of Linguistics and Applied Language Studies developed the score criteria for the Speaking Performance in much the same manner as for the other components of the CAEL Assessment. As with the Listening and Reading Performance, a raw score is initially calculated and is then converted into the standard band scores. This conversion is conducted on the basis of results from extensive pilot testing of test takers with a wide range of speaking proficiency.

The score criteria are shown in table 4.5 and encompass a broad range of speaking proficiency from the *Very Limited Speaker* (10-20) through to the *Expert Speaker* (80-90). This information is included in the CAEL Assessment Score Report provided to every test taker who has taken the Oral Language Test.

As described earlier, the Oral Language Test component of the CAEL Assessment is a tape-mediated process. The test takers responses for each item in the Oral Language Test are tape-recorded. The tape-recorded responses are then rated by a trained rater using a detailed (analytic) scoring criterion. The raw score obtained from this procedure is converted into a band score as for all components of the CAEL Assessment.

The inter-rater reliability of this procedure is monitored in essentially the same manner as for the other sub-tests. At regular intervals a selection of test taker response tapes are re-marked, the scores are converted to band scores, and a Spearman rank order correlation co-efficient is computed between the two sets of scores. The most recent correlation for a sample of n=178 oral language tests resulted in a correlation of .952. CAEL test developers are confident that the scoring methods used for the Oral Language Test produce reliable results.

Table 4.5: Speaking Performance Band Score Criteria

Band Score	Score Criteria
10-20	Very Limited Speaker speaks with great difficulty and many long pauses mispronounces many words manages to communicate some information
30	Limited Speaker speaks with some difficulty; hesitations or false starts mispronounces some words searches for words or provides studied and careful responses
40	Marginally Competent Speaker speaks with some fluency but without flexibility speed of response (either too fast or too slow) sometimes limits communication communicates information adequately but with noticeable effort
50	Competent but Limited Speaker speaks with some fluency and flexibility speaks unevenly – at times there is a natural and easy quality to the response and at other times the response breaks down
60	Competent Speaker speaks fluently, flexibly and with a degree of ease compensates strategically for limitations communicates most required information clearly
70	Adept Speaker speaks with ease presents information clearly and logically communicates required information effectively
80-90	Expert Speaker speaks with authority on a variety of topics demonstrates flexibility and controls nuance speaking is characterized by spontaneity and comprehensibility

Overall Result - Score Criteria & Methods

The Overall Result provided to CAEL test takers is neither a summation nor an average of the four sub-tests. Rather, a placement team meets to consider the entire score profile as well as factors such as the performance in specific sub-tests before assigning an Overall Result¹.

¹ Overall Results are closely tied to the proficiency levels of the ESL/EAP courses at Carleton University. The criteria of the scale map directly onto the continuum of proficiency which is denoted by course levels. In making decisions related to the overall score, the placement team considers which course would most closely match the test taker's English language learning needs. The link between language learning and proficiency is direct and contributes to validity arguments based on construct representation (Messick, 1989.)

The band scores for the Overall Result are listed in table 4.6 along with a description of the meaning of each Overall Result. This information is provided to test takers on the CAEL Assessment Score Report. The Overall Result criteria reflect the standards of proficiency which are accepted at Canadian universities who use the CAEL Assessment. Thus, the Overall Result criteria do not offer a description of English language performance such as is provided for the individual sub-test criteria.

Table 4.6: Overall Result Band Score Criteria

Band Score	Score Criteria
10-40	needs to increase the level of academic English before admission requirements for Canadian University degree programs are met
50	may meet academic English language requirements for admission to a few Canadian degree programs
60	meets academic English language requirements for admission to some Canadian University degree programs
70-80	meets academic English language requirements for admission to most Canadian University degree programs
80-90	meets academic English language requirements for admission to Canadian University degree programs

As indicated above, the overall score on the CAEL Assessment results from a review of all of the evidence collected about a test taker's ability to use English for academic purposes during the CAEL testing procedure. This includes:

- performance on the Oral Language Test,
- performance on the CAEL sub-tests of reading, listening and writing,
- responses to the personal essay and self-assessment, and
- other information which has been shown to have an effect on student performance in academic settings (e.g., amount of study in an English-medium school, intended program of study, intended level of academic study, etc.)

The CAEL Assessment review procedure is undertaken by the Testing Coordinator, the Coordinator of the Intensive English (ESL), the Academic English (EAP) Programs Coordinator and/or the CAEL Head Administrator. Because the review procedure involves judgement, the reliability of the procedure is constantly monitored. On a regular basis, 90-100 randomly selected and previously evaluated score summaries are re-evaluated by the CAEL Assessment review committee and a final score re-calculated. The correlation between the overall scores awarded at the first and second occasions is calculated.

Using a Spearman rank-order correlation co-efficient (ρ), the most recent re-evaluation of 178 score summaries results in a correlation of .948. The distribution of score changes for these 90 scores is shown in Table 4.7 below. The frequency of score changes is indicated as a percent. No change indicates that the band score assigned by the second review committee was the same as the band score assigned by the original committee.

A change of one indicates that the band score assigned by the second review committee was either one band score above or one band score below the original band score assigned by the first committee. There were no changes greater than one band score above or below the original band score. CAEL test developers are confident that the CAEL Assessment review procedure demonstrates adequate reliability.

Table 4.7: Overall Performance Band Score Changes (n=178)

Band Score difference	Frequency: Number and (%)	
No change	(n=148)	83.1%
Change of 1 band score	(n=29)	16.3%
Change of 2 band scores	(n=1)	0.6%

Comparability of Test Versions

For security reasons test versions are administered on a rotating basis. Although no two versions of any test may be considered perfectly comparable in all respects, versions of the CAEL Assessment are developed in a manner that meets recognized standards for comparability and satisfies principles of fair and ethical testing practice. Evidence of the comparability of the various versions of the CAEL Assessment is continually being gathered.

Evidence demonstrating comparability can be found from four sources:

- an examination of test development procedures;
- a statistical analysis of the score distributions for each version of the test;
- a statistical analysis of the scores of test takers who have taken more than one version of the test in a short time period; and
- specific research studies designed to investigate the comparability of the test versions.

Each of these sources of reliability evidence is described in this chapter.

In terms of test development, a number of procedures have been adopted in an effort to ensure the comparability of the test versions. First and foremost in this process was the development of test specifications for each component of the CAEL Assessment. These test specifications were developed through an iterative, consensus-based method involving both test developers at LATRU and experienced ESL/EAP teachers at the School of Linguistics and Applied Language Studies. The use of these test specifications maintains comparability across versions of the CAEL Assessment both in terms of the specific tasks contained in the test and in terms of the overall domain coverage of each sub-test.

Another test development procedure that helps to ensure comparability is the pilot testing process. Through extensive pilot testing, the performance of students with a wide range of proficiency on a test version under development can be compared with the performances of a similar group of students on an existing version of the test. The results of this comparison may be used to adjust individual test items that do not seem to be discriminating among test takers in the expected manner. As a final step, before releasing a new test version, test takers may be randomly assigned to complete one of two test versions and the resulting score distributions can be compared. Any difference in the performance of the two versions can be corrected before the new version is released for use at a test centre.

A statistical analysis of the score distributions is conducted on a version-by-version basis at regular intervals to ensure that all test versions function in a comparable manner. Given that there are currently 14 different versions of the test in use, it is not feasible to present all possible comparison distributions in this manual. However, the distributions can be compared for a subset of test versions. In Table 4.8, the Overall Result distributions are provided for four different versions of the CAEL Assessment.

These four versions were selected because, in every case, there were more than 400 recent Overall Result scores available. For reasons of test security, the versions are simply identified as A, B, C, and D. The Overall Results displayed in Table 4.8 represent CAEL test takers who completed the test at the main test centre at Carleton University in the period from January 2005 through September 2008.

Table 4.8: Frequency (%) of Overall Result Band Scores for Four Test Versions

Overall Result Band	Version A (n=434)	Version B (n=704)	Version C (n=762)	Version D (n=818)
10	1.4	2.8	1.6	2.3
20	16.1	16.9	15.9	14.3
30	30.0	26.4	30.7	34.5
40	27.6	24.0	28.5	27.8
50	15.2	18.3	12.1	8.9
60	5.3	7.8	7.1	6.4
70	4.4	3.3	3.9	5.0
80	-	-	.3	.9
90	-	-	-	-
Total	100	100	100	100

An examination of Table 4.8 reveals that, while there are some fluctuations in the band scores, the distributions for the Overall Result follow essentially the same pattern for each of the test versions.

Further evidence of the comparability of the different test versions can be found in an examination of the results of test takers who take two different versions of the test over a relatively short time period.

Apart from test takers who agree to participate in research studies, it is difficult to obtain data from actual test takers who have taken two different versions of the test. The majority of test takers only complete one version of the test. For those who complete two or more versions of the test there is often a long time delay and considerable language learning which occurs between the two testing occasions. However, we do have one source of data available to us from routine test administrations. Occasionally individual test takers make a request to the test administrator to re-write the CAEL Assessment. The test taker may feel that their performance on their initial test day was adversely affected by fatigue, illness, a recent arrival in Canada, or the particular version of the test which they wrote. Each of these requests is reviewed on an individual basis before permission to re-write the test is granted. If permission is granted to re-write the test then the test taker may complete a second CAEL Assessment. We track the performance of test takers who complete re-writes and the information which is gathered can shed some light on the reliability of the different test versions.

An analysis of the score results from those test takers who requested re-writes is reported in this chapter. The sample includes all test takers who wrote two different versions of the CAEL Assessment within a maximum period of four weeks. We have not included test takers for whom the interval is greater than four weeks because it is quite likely that a substantial amount of language learning could occur as the time interval between the test dates increases. For some test takers the interval between test dates is as short as one week. For others, an interval of two, three, or four weeks occurred. Table 4.9 offers a summary of the differences found in the Overall Result between the first and second testing occasion for these test takers (n=107). This data was collected over the period from January 1997 through to September 1999 and includes all of the test versions currently in use.

Table 4.9: Frequency of Band Score Changes in Overall Result

Overall Result - Band Score Change	Freq	%
2 band scores lower	6	5.6
1 band score lower	7	6.5
No change	47	43.9
1 band score higher	30	28.0
2 band scores higher	17	15.9
Total	107	100

It is clear from an examination of Table 4.9 that about 44% of test takers who take a second version of the test within a four-week time period score exactly the same in terms of their Overall Result. Further, 78.4% of the test takers have a final result within one band score of their original Overall Result. While 28% of the test takers score at one band score higher on the second testing experience, these results are essentially to be expected. There are three principal reasons for this modest gain in performance. First, all of the test takers in this sample were dissatisfied with their initial performance.

That is, these test takers requested a re-write because they felt that they had been adversely affected by factors such as fatigue or illness during their first test experience. This means that we can expect some improvement in their test scores because they were feeling better at the time of the second testing. The second factor which undoubtedly accounts for some of the improvement in these test scores is termed a 'practice effect'. That is, the test takers were much more familiar with the CAEL Assessment testing process. This familiarity is likely to result in a reduction in test anxiety and an improvement in test performance. The third factor which may account for the improvement in test scores is that some language learning has occurred between the first and second testing occasions. Very few test takers, only 6.5% found a reduction of one band score in their Overall Result on the second testing occasion. This analysis of test takers who have requested to write a second version of the test provides an additional source of evidence for the comparability of the different test versions currently in use.

More evidence of the reliability of the various versions of this test can be found in specific research studies which have been undertaken with the CAEL Assessment. An early study which compared the performance of test takers across two versions of the CAEL Assessment is reported in Zumbo, Fox, & Pychyl (1993). In this study the performance of 80 test takers who completed two versions of the test within a three week time period is examined. Measures of reliability, which are suitable for use in criterion-referenced testing situations, were computed. The results indicated that the decision consistency of the Overall Result for the CAEL Assessment was very high.

Jennings, Fox, & Shohamy (1999) investigated the potential existence of a topic-effect for the CAEL Assessment. In this study test takers (n=254) were randomly assigned to two groups. One group was given the opportunity to choose from among five versions of the CAEL Assessment. The other group was given the regularly scheduled version of the test. A detailed explanation of the experimental design and the analysis of the resulting data is available in the article (see Jennings et al, 1999). The results demonstrated that there was no effect of topic on the scores of the test takers. That is, the performance of the test takers that were given a choice of test version did not differ from that of those who had no choice. This evidence supports the contention made by CAEL test developers that the versions of the CAEL Assessment function in a comparable manner.