

# Validity

## Introduction

CAEL test developers have adopted an approach to validity that is consistent with the AERA/APA/NCME Standards (1985) and also reflects the writings of Samuel Messick (1989). Thus, CAEL test developers support the notion that "validity is an evolving property and validation is a continuing process" (Messick, 1989, p13). Through the ongoing process of test development and in-house research studies, evidence of the validity of the CAEL Assessment is continuously gathered. From time to time additional validity evidence is also provided in research studies, masters' theses and doctoral dissertations conducted at academic institutions outside the Language Assessment and Testing Research Unit (LATRU) at Carleton University. Readers interested in obtaining these documents may do so by contacting the Testing Coordinator (for full details see page one of this manual).

CAEL test developers embrace the definition of validity set out in the AERA/APA/NCME Standards document:

Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which that evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (1985, p. 9)

Traditionally, validity evidence has been gathered in three distinct categories: content-related, criterion-related, and construct-related evidence of validity. More recent writings on validity theory stress the importance of viewing validity as a 'unitary concept' (see Messick, 1989; AERA/APA/NCME Standards, 1985). CAEL test developers support the notion that validity is a unitary concept. Thus, while the validity evidence is presented in separate categories, this categorization is principally an organizational technique for the purposes of the presentation of research in this manual. The reader is encouraged to recognize that there is considerable overlap from one category of evidence to another. For example, a study which principally investigates criterion-related sources of evidence may also have real merit as a discussion of the construct-related validity of the test inferences. Finally, in keeping with the suggestion of Messick (1985), effort has also been expended to investigate the social consequences of the use of CAEL Assessment scores.

The evidence which has been gathered to date is presented in four categories: Ensuring Construct Representation, Investigating Construct Irrelevant Variance, Gathering Criterion-Related Evidence of Validity, and Consequences of Test Use.

## **Ensuring Construct Representation**

As identified by Messick, one of the principal threats to the validity of any test is that of construct under-representation. Construct under-representation occurs when "the test is too narrow and fails to include important dimensions or facets of the construct" (Messick, 1989, p.34). Great care was taken to address this issue when the test specifications were developed for the CAEL. The process began with an in-depth 'needs analysis' of the English language requirements of first-year students in a variety of faculties at Carleton University. The purpose of the needs analysis was to attempt to define the reading, writing, speaking and listening requirements of students in the academic setting. The needs analysis included surveys and interviews with professors and students as well as an examination of the assignment and exam methods used in courses throughout the university. Test developers attended lectures and reviewed the lists of articles and textbooks which students were assigned to read. The results of this analysis are described in greater detail in Fox, Pynchyl & Zumbo, 1993 and in Fox 1995). Among the most important observations revealed in the needs analysis was that the extensive use of multiple choice items was found to occur only in the departments of Psychology and Economics and then only in first-year courses. The remainder of departments placed more emphasis on the use of short-answer and essay responses. Even the department of Mathematics used some extensive writing as students were required to describe the process they used in solving problems.

This information has heavily influenced the construct of English for academic purposes which is measured in the CAEL Assessment. The results of the needs analysis were the principal reason for adopting a constructed-response question format in the CAEL Assessment. Item types including filling in the blanks, labeling of diagrams, completing tables, note-taking, short answer and longer essays are used in the CAEL Assessment because these tasks are common within the Canadian university context.

More detailed test and item specifications were developed when the results of the needs analysis were complete. Each detailed CAEL Assessment specification includes the domain, title and general description of the criterion. A prompt or task definition is included to explain exactly what the student will be required to do for each test item. Item specifications also provide a response definition explaining what the student will do in responding to the prompt. This includes a description of successful and unsuccessful responses. A sample item or task is included with every item specification and additional information defining the characteristics of the context of use is provided. The use of these detailed test and item specifications in the development of new versions of the CAEL Assessment helps to ensure that the construct validity of the test remains true to the original concept of the test.

Follow-up studies are conducted at regular intervals to determine the extent to which faculty members at the university find the language requirements of the CAEL test to be consistent with the English language use in their courses.

## Investigating Construct-Irrelevant Variance

The second principal threat to the validity of any test is the potential for construct-irrelevant variance. Construct-irrelevant variance exists when the "test contains excess reliable variance that is irrelevant to the interpreted construct" (Messick, 1989, p.34). This construct-irrelevant variance is viewed as a contaminant with respect to score interpretation. For this reason, CAEL test-developers attempt to investigate and minimize all sources of construct-irrelevant variance. Two kinds of construct-irrelevant variance can be identified. Construct-irrelevant difficulty occurs when "aspects of the task that are extraneous to the focal construct make the test irrelevantly more difficult for some individuals or groups" (Messick, 1989, p. 34). Conversely, construct-irrelevant easiness occurs when "extraneous clues in item or test formats permit some individuals to respond correctly in ways irrelevant to the construct being assessed" (Messick, 1989, p. 34). In general then, construct-irrelevant difficulty leads to lower scores for some test takers while construct-irrelevant easiness leads to higher scores for some test takers. CAEL test developers are constantly examining the test specifications, test items, administration and scoring techniques in an effort to reduce the impact of these threats to validity.

Efforts to investigate sources of construct-irrelevant variance generally take the form of research studies which are presented at academic conferences, peer reviewed and published. This process enables test developers to scrutinize various aspects of the test and to receive critical reviews from professional language testers throughout the world.

Probably the single greatest potential source of construct-irrelevant variance for the CAEL Assessment results from the integrated topic-based nature of the written test (the Oral Language Test is task-based). Because each version of the written test discusses one specific topic, it is possible that some test takers may be advantaged or disadvantaged in terms of their performance because of the topic of the test they wrote. That is, individual test taker's background knowledge, interest and opinions concerning a topic may impact their performance. The test specifications for the CAEL Assessment recognize this potential source of construct-irrelevant variance. For this reason, care is taken to ensure that all the information the test taker requires to formulate responses is contained within the testing materials provided. The content of the readings and the lecture provide a rich context, which test takers can draw upon in making their responses. Attempts to build the type of *context-free* items which are often claimed by other English language proficiency test developers are not seen as appropriate for the construct of English for academic purposes which the CAEL Assessment endeavours to measure.

Nonetheless, CAEL test developers take seriously the responsibility to demonstrate that the topic of the test is not a source of construct-irrelevant variance. For this reason, studies have been conducted and published investigating this potential validity threat. The most extensive study of this issue is described in Jennings, Fox, Graves & Shohamy (1999) and has already been described in some detail in the '*Comparability of Test Versions*' section of Chapter Four, page 55.

The results of this study which involved a comparison of test taker performance across six different versions of the test found no impact of topic on the scores of test takers (n=254). CAEL test developers are confident that the topic of the test is not a source of construct-irrelevant variance.

### **Gathering Criterion-Related Evidence of Validity**

Gathering criterion-related evidence of validity is an important task for all language testers. This task is particularly difficult for the CAEL test given some of the unique features of the CAEL Assessment. That is, the use of constructed-response test items in a topic-based fully integrated language test is essentially a unique approach to language testing at the present time. These aspects of the CAEL Assessment strengthen the claim made by test developers that the CAEL Assessment is a reasonable approximation of the language demands of English for academic purposes, particularly in Canadian university contexts. However, the essentially unique nature of the test means that gathering criterion-related evidence of validity is problematic.

CAEL test scores have been compared with the performance of test takers on the Test of English as a Foreign Language (TOEFL). However, the TOEFL is clearly measuring English language proficiency in a very different manner. Does this mean, then, that a correlation of the two test scores provides criterion-related evidence in support of the CAEL test? Clearly, the establishment of an appropriate criterion is always a challenge when gathering this type of validity evidence.

One procedure that has been adopted in an effort to gather more meaningful criterion-related evidence of validity is to conduct follow-up studies of CAEL test takers who score at various proficiency levels. One such follow-up study was conducted for this manual. In this study the university course performance of 79 test takers who achieved an Overall Result at a band score of 70 or greater was collected. The basic design was to determine the grade point averages (GPA) of these students in their first full term of study after achieving an Overall Result of 70 or greater on the CAEL Assessment. The score of 70 was selected for this study because test takers who achieve this score are permitted to register for regular courses at the university without any further ESL/EAP training. Data was collected for each test taker for the term immediately following their CAEL Assessment in an effort to avoid measuring the impact of language learning which occurred after the test was completed.

A six-point scale was used for the GPA as shown in Table 6.1.

**Table 6.1: Grade Point Average and Letter Grade Equivalents**

<b>Letter Grade</b>	<b>Grade Point Value</b>	<b>Letter Grade</b>	<b>Grade Point Value</b>
A+	6.0	C+	3.0
A	5.5	C	2.5
A-	5.0	C-	2.0
B+	4.5	D+	1.5
B	4.0	D	1.0
B-	3.5	D-	.5
		F	0

In the sample of 79 test takers the range of GPA's was from 0 through to 6.0. The frequency distribution of GPA's for this sample is shown in Table 6.2.

**Table 6.2: Frequency Distribution of Grade Point Averages**

<b>Grade Point Average Range</b>	<b>Freq</b>	<b>%</b>
0 to 0.49 (F)	2	2.6
0.5 to 1.9 (D +/-)	6	7.7
2.0 to 3.4 (C +/-)	15	19.1
3.5 to 4.9 (B +/-)	28	35.4
5.0 to 6.0 (A +/-)	28	35.4
Total	79	100

We can begin to interpret this information in absolute terms. It would appear that the vast majority of test takers who took the CAEL Assessment and were permitted to register in university courses as a result of their strong performance have been able to successfully complete their courses. The mean GPA for the entire sample was 4.11. This indicates a GPA at the letter grade level of a B. While table 6.2 indicates that 2 of the 79 students achieved a GPA equivalent to an F, it is very difficult to determine the extent to which their difficulties were related to English language proficiency. However, it is clear that the vast majority of the students are performing in an acceptable manner after completing the CAEL test.

### **Consequences of Test Use**

Messick (1989), among others, highlights the importance of considering both the intended and unintended consequences of test use when accumulating evidence of the validity of inferences for a given test. This emphasis on the consequences of test use is also included in the principles of fair and ethical testing described in the Standards document referred to earlier in this manual (AERA, APA, & NCME, 1985). CAEL test developers are committed to ensuring that the information provided to test takers and the institutions which use CAEL Assessment Score Reports is a consistent and meaningful measure of the candidates' proficiency in the use of English for academic purposes.

The use of the CAEL Assessment in lieu of other currently available measures of English language proficiency arguably provides the test taker with a more accurate measure of the construct of English for academic purposes. However, it is important that the test takers and test score users have the same confidence in the validity of the inferences made from CAEL Assessment results as is held by the test development team. In an effort to determine the extent to which test takers and test score users share this impression of the CAEL Assessment, feedback is solicited through a number of means.

CAEL test developers gather on an on-going basis the feedback of test takers, score users and members of the university community who are impacted by the CAEL Assessment. Efforts to gather and review this feedback help to minimize any potentially negative consequences of the use of CAEL Assessment scores.

### **Test Taker Feedback: Post-Test Questionnaire**

At the end of every CAEL test, test takers are asked to complete a one-page questionnaire. The questionnaire is designed to allow test developers and administrators to gather the comments and opinions of test takers concerning the CAEL Assessment. At present, there are three types of questions on the questionnaire. In the first type of question, test takers are given statements concerning the CAEL Assessment and asked to respond by circling agree, disagree or no opinion. The second question asks test takers to rank five factors in terms of which factor they find most important in their test experience. Finally, the questionnaire includes three open-ended items in which test takers are given the opportunity to indicate what they like about the test, what they dislike about the test, and what aspects of the test might be changed in order to improve their performance. The test takers' responses to the questionnaire are summarized at regular intervals and are considered carefully when tests are being developed and when changes are being made to the test administration process.

In response to the statement 'This test was a good experience', 83.2% of test takers agree, 5.6% disagree, 8.4% indicate that they have no opinion and 2.9% did not respond to the item.

The responses from a recent summary of 790 test taker questionnaires collected from the period from January 1997 through September 1999 is presented in this chapter to illustrate the type of information which has been collected. In response to the first 'agree/disagree' statement 'This test reflects my true knowledge of English.', 39.1% of test takers agree, 32.0% disagree, 23.8% indicate that they have no opinion and 5.1% did not respond to the item. While the largest group of test takers indicates that they feel the test does reflect their knowledge of English there is still a relatively large proportion of test takers who disagree with the statement. Some further insight may be gained from the open-ended items on the questionnaire. Some test takers felt that they lacked sufficient information to assess their own knowledge of English. These test takers tended to respond in the 'no opinion' category. Other test takers clearly felt that their performance on the test was not as strong as it could be and that their knowledge of English was greater than that revealed by their test performance.

These test takers tended to disagree with the statement. It is hoped that continued research will provide evidence as to the extent to which this tendency to feel the test does not reflect 'true knowledge' is an artifact of all testing settings and the extent to which it is specific to the CAEL Assessment experience.

In response to the statement, "This test was fair," 56.1% agree, 13.7% disagree, 24.7% indicate that they have no opinion and 5.6% did not respond to the item. Clearly, while some test takers feel that the test did not reflect their true knowledge of English, most test takers feel that the CAEL Assessment was fair.

In the second question, test takers are asked to rank the importance of five factors in their test performance. The five factors are the amount of time allowed to complete test items, the physical comfort of the testing environment, the availability of sample tests, the sound quality of the taped-lecture, and the topic of the test which the test taker completed. These factors have been identified by test takers in response to the open-ended items of the questionnaire and have also been reported on in a number of research studies.

A recent summary of 790 test taker questionnaires indicated that 29.5% of the test takers did not answer the ranking question. Further analysis of these results indicated that over 50% of those test takers who did not answer the item had achieved Overall Results with band scores at 40 or lower. This may indicate that the format of the item is too difficult for test takers with lower levels of English language proficiency. For this reason, the pattern of responses with respect to this item more closely reflect the opinions of test takers with intermediate or advanced levels of English language proficiency as assessed by their CAEL performance. The factor that was most frequently identified as the most important by these test takers was the amount of time allowed to complete the test items. In fact 25.9% of test takers indicated that this was the most important factor in their test score performance. This finding is consistent with the results of other research studies which examined the test takers' responses to testing conditions (Norton & Starfield, 1997). The factor which was identified the second most frequently as the most important factor in the test takers testing experience was the topic of the test. From this sample 22.7% of the test takers felt that the topic of the test was the most important factor in their performance.

The responses to the open-ended sections of the post-test questionnaire seem to support the observations from the ranking item. In the three open-ended items on the questionnaire test takers are asked to indicate what they liked about the test, what they disliked about the test and what changes to the test would help them most. It is difficult to summarize the results from these open-ended questions in a manner which is brief enough for inclusion in the test manual. However, the essential findings were that the majority of open-ended comments re-iterated a need for more time to answer the items. The second most frequently appearing comment was that the topic of the test was problematic.

## Test Impact over Time

CAEL test scores have important consequences for test takers, as is the case with all high-stakes language tests. Achieving a certain score on a CAEL Assessment may mean the difference between being accepted to study in a university program or being denied admission. Therefore, it is important to examine the adequacy of CAEL test decisions through the on-going collection of evidence regarding test takers whose lives are affected by their performance on the CAEL Assessment.

In order to examine the relationship between CAEL test performance, which occurs within the highly constraining context of a testing setting, and actual language ability demonstrated in a classroom setting over time, data are collected on an on-going basis within the English for Academic Purposes program at Carleton University. Within the Carleton context, a CAEL test taker's overall band score is related to a level of English language ability that is linked to and reflective of one of a range of EAP courses offered by the School of Linguistics and Applied Language Studies<sup>1</sup>. Thus a band score of 40 allows a test taker to register for a specific, introductory course in English for Academic Purposes and take one additional course in their discipline. A band score of 50 allows a test taker for a specific, intermediate course in English for Academic Purposes and to take two additional courses in their academic discipline. The organization of the EAP program and its relationship to performance on the CAEL Assessment creates a context wherein the adequacy and usefulness of the CAEL scores may be examined over time.

For example, a 1999 study of test impact utilizes the overall CAEL Assessment scores of 120 test takers who wrote the CAEL Assessment during July and August, 1999 at Carleton University and were later registered in one of the 12-week EAP courses for the Fall term (September-December, 1999). At the end of ten weeks of instruction, a questionnaire was circulated to the 12 teachers who received these students. The questionnaires identified the test takers in their classes who had been placed there on the basis of the test<sup>2</sup>, and asked the teacher's evaluation over time. Nine (9) of the teachers responded regarding the performance of ninety five (95) students. The correlation between the level of language ability demonstrated by the test taker performance on the CAEL Assessment and the language ability demonstrated by the student in class is significant ( $<.01$ ), with a correlation co-efficient of .863 based on a Spearman rank-order correlation.

---

<sup>1</sup> As is the case in a number of other universities, Carleton has always been described as having a "gradual admission" policy. Students are allowed to begin their university program provided they include a required EAP course during the first term/s of study and a limited number of courses in their chosen discipline, according to a formula defined by the School of Linguistics and Applied Language Studies. Thus, students testing at band score 60 on the CAEL Assessment would be required during their first term of study to register in ESLA 1900, an advanced EAP course, and allowed to take three other courses specific to their programs of study. The students would earn credits toward their degree from the EAP course as well as the other academic courses.

<sup>2</sup> Although all students test into an EAP course on the basis of a CAEL test score, they may satisfy the language requirement by successfully completing the highest EAP course, ESLA 1900.

In addition to Spearman’s rho, the distribution of differences between CAEL test performances and the teachers’ evaluations for these 95 test takers is reported below in Table 6.3. No change indicates that the band score on the CAEL Assessment was, in the teachers’ opinions, an accurate indication of the students’ actual English language ability as demonstrated by their in-class performance during the 10 week period. A change of one either indicates the CAEL test performance under- or over-estimated the students’ English language ability.

**Table 6.3: Agreement between CAEL Assessment Placement and Teacher Evaluations (n=94)**

<b>Grade Point Average Range</b>	<b>Freq</b>	<b>%</b>
No change	76	81%
Within One/half to One Band	15	16%
Within Two Bands	3	3%
<b>Total</b>	<b>94*</b>	<b>100</b>

\*One test taker withdrew during the 10-week period.

As table 6.3 indicates, there is considerable agreement between the test taker’s performance on the CAEL Assessment and the EAP teachers’ evaluation of ability to use English based on 10 weeks of evaluation in the context of the classroom.