Commission de la fonction publique du Canada
Public Service Commission of Canada

Testing in the Public Service of Canada

Standards for the Development and Use of Tests for Appointment Purposes

Date: *September, 2007*
Author: *Assessment Oversight*
Policy Development Directorate

Canada

TABLE OF CONTENTS

INTRODUCTION

Standardized tests, whether administered to assess cognitive abilities, aptitudes, interests, skills, or personality, represent one of the most important contributions of behavioural science to our society. A wealth of evidence tells us that the proper use of well-constructed tests provides a sound basis for making important human resource management and personnel appointment decisions.

In an employment context, managers need accurate and reliable information on how competently applicants can be expected to perform on the job. Tests can provide much of this information. A test consists of a set of questions or tasks that are presented to applicants under standardized conditions, and they respond in ways that demonstrate the competencies required for the job. The results of well-developed and properly-applied tests can be used to evaluate how qualified a particular applicant is for a job relative to others and predict his or her job performance.

Tests come in a variety of forms. For many people, the word "test" evokes the traditional paper-and-pencil test in multiple-choice, short answer or essay format. However, many other selection instruments may be called "tests". These include everything from job sample instruments, such as written in-basket exercises and interactive job simulations, to structured interviews and assessment centres (which normally consist of a number of "tests"). They can be written or administered face-to-face, but also through other means, such as computer or telephone. In this document, the generic term "tests" is used to refer to all of the assessment instruments that are used in personnel decision-making practices.

This document is organized in two parts: Part A contains Standards for well developed and administered tests. The evaluation of the appropriateness of a test or application depends on judgement on the part of the user. The Standards provides a frame of reference to ensure that relevant questions are examined when choosing a particular test.

Part B outlines the legal basis and policies that regulate test use, to which test users must conform, in the public service of Canada.

PART A: STANDARDS FOR TESTING IN THE PUBLIC SERVICE

The purpose of this part of the document is to present essential Standards for the development and application of tests used for the appointment of personnel in the Public Service at a level that will be useful for managers and human resources practitioners.

The thrust of the Standards is that a sound technical and professional basis, derived from research, must underlie the development and use of tests. S*tandards for Educational and Psychological Testing (1999)* by the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education were used as a basis for the development of the Standards.

Sections 1 to 5 contain Standards concerning test instrumentation: validity and reliability; test development; test scoring, equating and related issues; and test documentation. Sections 6 and 7 contain Standards for test use and administration, including testing persons with disabilities.

Informed judgement must be used to determine which of the Standards is relevant, and in which ways, for a particular test or application. The evaluation of a test or application does not rest on the satisfaction of every Standard, and cannot be determined using a checklist. Not all of the Standards apply to all types of tests or to all applications of a test. This is why determining the relevant standard to be respected will depend on several factors. These include the informed judgement of the test user in evaluating the type of test, the particular circumstance, the purpose for which the test is being used, and the availability of information supporting the evaluation. All these elements will play a part in determining the appropriateness of a Standard.

The Standards are not meant to prescribe the use of certain statistical methods rather than others. For example, where specific methods are mentioned, it is understood that these are related to current professional standards which are by themselves subject to change over time. Where specific statistical reporting requirements are mentioned, the phrase "or equivalent" should be understood.

The Standards were prepared with full awareness of the concerns over the impact of testing on designated employment equity groups. The Standards should be applied with due regard for designated groups, and one section is devoted to the evaluation of applicants with disabilities.

1. Validity

Validity is the single most important consideration in developing and evaluating tests. In the context of personnel selection, the test user draws inferences about job or training performance from applicants' scores on a test. These inferences are said to be *valid* when there is enough evidence to support the claim that there is a substantive relationship between scores on the test and some aspect of work behaviour. Such behaviours typically include the quality of job performance and performance in training. They may also include advancement, termination, or other organizationally pertinent behaviour.

Informed judgement should guide decisions regarding the sufficiency of the evidence supporting claims of validity. V*alidation* is the process of gathering the evidence which is required to support the claim that test scores are related to job performance. This rationale is referred to as the *validity argument*.

Standard 1.1

Validity evidence should exist for all intended usage of a test. It is assumed that, for each use contemplated, there is a need to determine the validity of a given test.

Standard 1.2

The demonstration of validity is a joint responsibility of the test developer and the test user. Test developers are responsible for providing all available information which may assist in determining the validity of a given test developed for a specific intended purpose. Test users have the responsibility of ensuring that qualifications or competencies for which they use a test are relevant to the job and that the test assesses these qualifications.

Gathering Validity Evidence

There are various types of evidence that can be gathered to demonstrate a test's validity. While different strategies for gathering such evidence may be used, the primary inference in the employment context is that a score on a selection procedure predicts subsequent work behaviour.

Standard 1.3

Evidence should be provided supporting the link between test scores and job performance. Studies or other published data which are used as evidence should be cited appropriately, and other empirical evidence or unpublished data should be submitted as necessary.

Standards 1.4 through 1.6 address the three sources of evidence that can be used to support the validity argument:

a) Evidence based on relationship with other variables.

This type of evidence is based on the relationship between test scores and external variables. It involves examining the relationship of test scores to other variables that measure the same thing (*convergent evidence)* or different things (*discriminant evidence).* This can be achieved through two general approaches: *predictive* and *concurrent.* A predictive study looks at how accurately test scores can predict criterion scores (i.e. future job performance) that are obtained at a later time while the concurrent study obtains predictor (test) and criterion data at the same time.

When developing a new test, sufficient data cannot always be collected to support validity evidences from the hiring organization. If this is the case, it may be possible to use *validity transportability* or *validity generalization.* Validity transportability refers to the situation where predictive relations from another specific setting are judged to be relevant to the current setting. Validity generalization refers to a situation where past validation studies in similar situations are used to provide a valid estimate of the relationship between test scores and job performance in the situation at hand.

Standard 1.4

When a validity argument makes use of analyses of the relationships between test scores and other variables, the evidence should be described in ways that are sufficient for others to understand its nature. For instance, if the aim is to relate test scores with job performance criteria, then it should be made clear whether the predictive or concurrent approach is being used.

b) Content-related evidence.

When the validity argument is based on test content the analysis focuses on the relevance and the representativeness of the test content with respect to the job or certain job functions. These functions in which the applicant will have to perform, are commonly called performance domains. In these cases, tests are understood as samples of performance domains. Often, the relevance and representativeness of the test content is determined by the judgements of subject matter experts, such as managers or job incumbents.

Standard 1.5

When a validity argument is based on analyses of test content, the procedures followed in generating the test content, the procedures for selecting subject matter experts and the kinds of judgements these subjects matter experts were called upon to make should be adequately described. For job selection and promotion, the definition of the performance domain should be based on an analysis of job characteristics relevant to the proposed test use.

c) Evidence based on response processes.

When a validity argument uses evidence based on response processes used by applicants when they take the test, it is necessary to demonstrate that the processes used are similar to those required in the targeted job. For example, if a claim is made that a work sample measures use of proper techniques for resolving customer service problems, then simply assessing whether the problem is resolved is not enough. It is necessary to clarify both mental and physical response processes which contribute to solving problems. Often evidence regarding individual responses can be gathered by a) questioning test takers about their response strategies, b) analysing examinee response times on computerized assessments, or c) conducting experimental studies where the response set is manipulated.

Standard 1.6

When a validity argument is based on evidence with respect to the processes deemed to be used by respondents in responding to the test materials, a conceptual model based on evidence linking observable events to the underlying process should be presented.

Interpreting Validity Coefficients

Validity is expressed as a number between 1.0 and -1.0. A value of 1.0 means that there is a perfect positive relationship between the score received on a test and performance on the job. A value of 0 means that there is no relationship, in practical terms, that the test does not predict job performance. A negative value indicates an inverse relationship: the better the performance on the test, the worse the expected on-the-job performance. It is very unusual for a validity coefficient to rise above 0.60 which is far from perfect prediction.

Test Validity and Fairness

The fairness of a test is related to its validity. In the context of personnel selection, a test is considered fair if it is a valid measure of competencies required for job performance and can be judged not to contain elements unrelated to those competencies, which would be responded to differently by different groups of test takers. At times, differences in group test performance may be observed. These differences by themselves say little about the validity or fairness of a test since these differences may reflect differences in applicant characteristics which influence test performance (e.g., amount and type of education).

2. Reliability

Reliability refers to how dependably or consistently a test measures a characteristic. A test taker may perform differently on one occasion than on another for reasons unrelated to the purpose of measurement. Differences between scores from one test form to another or from one occasion to another are attributable to what is commonly called errors of measurement. Such errors reduce the reliability of the score obtained for a person from a single measurement. Fundamental to the proper evaluation of a test is the identification of the main sources of measurement error.

Factors Influencing Reliability

Examples of sources of measurement error may include, but are not limited to, the following:

- **Test taker's psychological or physical state.** Test performance can be influenced by a person's psychological or physical state at the time of testing. For example, levels of anxiety, fatigue, or motivation may affect the applicant's test results.

- **Environmental factors.** Differences in the testing environment, such as room temperature, lighting, noise, or even the test administrator, can influence an individual's test performance.

- **Test form.** Many tests have more than one version or form. Questions differ on each form, but each form is supposed to measure the same thing. Different forms of a test are known as parallel forms or alternate forms. These forms are designed to have similar measurement characteristics, but they contain different items. Because the forms are not exactly the same, a test taker might perform better on one form than on another.

- **Multiple raters.** In certain tests, scoring requires judgement on the part of the rater. Differences in training, experience, and frame of reference between raters can produce different test scores for the test taker.

Types of Reliability Estimates

There are several types of reliability estimates each influenced by different sources of measurement error. Test developers have the responsibility of reporting the reliability estimates that are relevant for a particular test. The acceptable level of reliability will differ depending on the type of test and the reliability estimate used. Traditionally, four broad categories of reliability estimates have been recognized:

- **Test-retest reliability.** In this approach, the repeatability of test scores is evaluated by administering the same test at different times. This estimate also reflects the stability of the characteristic or construct being measured by the test. Some characteristics are more stable than others. For example, an individual's reading ability is more stable over a particular period of time than that individual's anxiety level. Therefore, you would expect a higher test-retest reliability coefficient on a reading test than you would on a test that

measures anxiety.

- **Alternate or parallel form reliability.** This approach evaluates how consistent test scores are likely to be if a person takes two or more forms of a test. For instance, a high parallel form reliability coefficient indicates that the different forms of the test are very similar, which means that it makes virtually no difference which version of the test a person takes. On the other hand, a low parallel form reliability coefficient suggests that the different forms are not comparable; they may be measuring different things and therefore cannot be used interchangeably.

- **Inter-rater reliability.** This approach evaluates how consistent test scores are likely to be if the test is scored by two or more raters. For example, there are some tests for which raters evaluate responses to questions and determine the score. Differences in judgements among raters are likely to produce variations in test scores. A high inter-rater reliability coefficient indicates that the judgement process is stable and the resulting scores are reliable. Inter-rater reliability coefficients are typically lower than other types of reliability estimates. However, it is possible to improve these reliabilities if raters are appropriately trained and given adequate feedback.

- **Internal consistency reliability.** This approach evaluates the extent to which questions on a test measure the same thing. A high internal consistency reliability coefficient for a test indicates that the questions on the test measure the same characteristic. It is important to note that the length of a test can affect internal consistency reliability. For example, a very lengthy test can spuriously inflate the reliability coefficient.

The Role of Test Developers and Users

The degree to which test scores are unaffected by measurement error indicates the reliability of the test. As such, test developers and users have important roles to play in maintaining test reliability. The test developer should indicate those aspects of test administration and scoring which are crucial to maintain the reliability of a test. Similarly, test users should ensure the standardized administration of the test and scoring of test responses.

Interpreting reliability coefficients

The degree of reliability of test scores is usually expressed in terms of a *reliability coefficient.* It is denoted by the letter "r," and is expressed as a number ranging from 0 to 1.00 with r=0 indicating no reliability, and r=1.00 indicating perfect reliability. Do not expect to find a test with perfect reliability. General guidelines for interpreting test reliability suggest that a coefficient of less than .70 has limited applicability; .70-.79 adequate reliability; .80-.89 good reliability, and .90 and higher excellent reliability. With these guidelines in mind, test users should not accept or reject a test solely based on the size of its reliability coefficient. To evaluate a test's reliability, you should consider the type of test, the type of reliability estimate reported, and the context in which the test will be used. This brings us to the reliability standards.

Standard 2.1

Information relevant to the reliability of test scores should be provided in adequate detail to permit decisions to be made about the usefulness of a test for a given application.

Standard 2.2

Information regarding the specific method(s) of estimating test reliability should be reported and the conditions under which the reliability estimates were obtained should be clearly described.

Standard 2.3

Test *users* should engage in procedures which maintain or enhance the reliability of test scores. Test *developers* should clearly indicate which aspects (e.g., scoring procedures, administration protocols) are essential to maintaining the reliability of test scores.

3. Test Development and Revision

The following Standards concerning test development apply to tests currently in use, to tests under revision, and to tests that are to be constructed. These Standards also cover how the content of the overall test and of the test questions are determined, how test questions are evaluated and ultimately selected for the test, and how the test design matches its intended use. Test developers are responsible for periodically reviewing their tests. As appropriate, the continued relevance of test content should be reaffirmed.

Standard 3.1

Tests programs should be developed on a systematic rational basis. The test developer should review the evidence relevant to a test and decide what evidence must be available before the test can be used and what can reasonably be provided later.

Standard 3.2

Test specifications should include the following: a statement of what the test is intended to measure, the number and format of the questions and their desired psychometric properties. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring.

Standard 3.3

Test developers should help ensure that their tests are fair to all test takers. They should avoid test content that could be considered offensive. Also, they should take into account the representation of designated groups in the intended examinee population, which should be reflected in samples that are used to construct the test. Test instructions should provide for fair and impartial treatment of test takers. Monitoring of operational use will often be appropriate when it is suspected that there is differential performance on the test by groups of test takers. If monitoring does reveal the presence of differential performance, additional studies should be carried out to better understand the nature of such differences.

Standard 3.4

The instructions to test administrators should provide sufficient guidance to ensure that the test is administered correctly. The directions prescribed for presentation to test takers should be complete so that test takers can be expected to respond to the task as the test developer intended. When appropriate, sample materials and/or sample questions should be provided for preparation.

Standard 3.5

When a test is made available in both official languages, the equivalency of forms in both official languages should be assured. The language used in a test should not be unduly difficult. The level of difficulty should reflect the level used on the job or group of jobs for which the test is designed.

Standard 3.6

Procedures for the scoring of tests should be presented by the test developers in sufficient detail and clarity to ensure the accuracy of scoring and interpretation. The administration materials and/or score report forms should be designed to facilitate appropriate interpretation of test scores. Where judgemental processes enter into test scoring, the bases for scoring and/or the procedures for training scorers should be presented by test developers in sufficient detail to permit accurate application of judgement. Where appropriate, the test developer might require specific training or certification for test users, administrators, scorers, and interpreters. Where normative interpretation is involved, it is the responsibility of the test developer to provide sufficient information on the normative group(s) for the test user to judge the relevance of such norms.

Standard 3.7

At times, accommodating the needs of persons with disabilities will require the modification of test materials or test administration procedures. The instructions to test administrators should then be clear as to how far standardized test materials can be modified and/or standardized test administration procedures can be modified without compromising its validity. When it is judged that the modification of standardized test materials and/or administration procedures would render test scores meaningless, an alternative approach to the assessment of the same construct should be considered. Testing accommodations involving the modification of Public Service Commission' standardized tests developed by the Personnel Psychology Centre (PPC) or its administration procedure can only be determined under PPC guidance.

4. Interpretation of test scores

The following Standards concern the interpretation of test scores. The topics addressed are as follows:

- Establishing a context;
- Establishing norms;
- Comparing scores from different test forms; and
- Determining cut scores or pass marks for tests.

Establishing a Context

The simplest level of information that comes from a test is a raw score, which for many tests is a count of the number of test questions answered correctly. Score interpretation is often aided by converting raw scores to some other scale of measurement. An example of such scaled scores (or derived scores) in government might be the A, B, and C designations assigned to levels of second language proficiency.

Test scores must take their meaning from some context. Sometimes scores can be directly interpreted by reference to a specific set of behaviours. For example, a sufficiently high score on a test of verbal language proficiency could indicate that the person can carry on a basic conversation in the language assessed. A test that generates such scores is called a criterion-referenced test. These scores reflect directly what the test taker can do.

Standard 4.1

Scales used for reporting scores and the rationale for their choice should be precisely described to facilitate accurate interpretation of scores by tests users. How scaled scores are derived from raw scores should be documented.

Establishing Test Norms

In order to meaningfully interpret test results, information is required on how relevant others have performed on the same procedure. This consists in comparing test results to overall results representing the performance of a group of people to which the test is usually administered. For example, a person applying for a Public Service position takes a numerical reasoning test. He or she obtains a raw score of 48 correct responses out of a possible 68. Is this score, average, above average, or below average? In and of itself, the score of 48 does not give much information. In order to interpret this score meaningfully, we need to compare the applicant's raw score to the distribution of scores of relevant others- that is persons of approximately the same occupational category and educational background who were being tested for the same purpose. These persons make up a *norm group*. The statistics that describe the performance of this group would be called *norms*.

Norms constitute a frame of reference for interpreting any one test score. Those responsible for determining which group should constitute the normative group must have adequate knowledge of the test and its purpose.

Standard 4.2

Norms must be relevant to the test takers to whom the test is usually applied. Enough information about the sample(s) on which the norms were collected must be documented so that the appropriate application of the norms can be determined. The adequacy of a set of test norms should be evaluated periodically for its continued relevance.

Comparing Scores From Different Test Forms

It is useful for a test to have multiple forms. For example, when someone who has recently been tested with a particular test is to be retested, it is desirable to use a second form of the test which contains similar, though not identical questions. The use of the alternate version helps to ensure that the person who is being retested does not gain an unfair advantage relative to those who have been tested only once. However, when two or more forms are used interchangeably it is essential that they are equally difficult. Ideally, the test taker should receive the same test score regardless of which form of the test was used. Another approach to obtaining equivalent scores is to use a statistical equating method. In these methods, scores on the test forms are converted into an equivalent scale. Many equating methods involve the use of norms.

In addition, alternate test forms may be needed for testing persons with a disability. These tests are designed so that persons with a disability have an equitable opportunity to demonstrate the competency being assessed. Versions for persons with disabilities are often presented in a different format (e.g., Braille or audiotape) and the administration may differ (e.g., only one applicant at a time).

Standard 4.3

When it is alleged that alternate forms of a test can be used interchangeably, documentation supporting the equivalence of the forms should be available. The specific methods for equating tests and for converting scores should be sufficiently described and documented.

Setting Pass Marks or Cut Scores

A test may have a cut score, or several cut scores, associated with it. Cut scores take a range of test scores and divide the range into categories, for example, "pass versus fail" and "most successful to least successful". Cut scores may be set by various authorized people, including test developers and users. Such people need adequate knowledge of the test and of the purpose for which the test will serve.

Standard 4.4

When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing them should be clearly documented. Cut scores may be determined using a variety of methods, including statistical data, judgement, and performance criteria. Whatever method is used, it is important to ensure that the cut score is relevant to the job competencies that the test is intended to measure.

5. Test Documentation

Information regarding test development, test use and test score interpretation should be available. Although often selected in consultation with testing experts, it is important to support test users to ensure their optimal use and to avoid errors in the application, administration and scoring. Tests are likely to be used by people with minimal training in testing. Such users need test information communicated to them with an appropriate level of detail. Other test users may be measurement specialists who need access to technical or statistical information to help them judge the adequacy of a test. Separate documents (e.g., candidate's brochure, administration manual, etc.) may be written to meet the needs of different test users, including test administrators, job applicants, researchers and so on.

Standard 5.1

Test documents (e.g., tests manuals, technical manuals, user's guides) should be made available to tests users.

Standard 5.2

Test documentation should describe the rationale for the test, recommended uses of the test, support for such use, and information to assist in score interpretation. Where particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.

Standard 5.3

Sufficient documentation should be available to support the reliability and validity of the test. The procedures used to estimate reliability and standard error of measurement and the size and type of samples involved should be described. Similarly, the validation sample and any limitations to the generalization of findings derived from it should be described. Evidence of validity for total test scores should be supplemented with documented evidence for other types of scores if scores other than total scores are recommended for use.

Standard 5.4

The test developer should have documentation available that describes changes that may have been made to a test over time.

Standard 5.5

Test documentation should identify any specific qualifications required of test administrators, such as language competencies, knowledge of certain types of testing materials, etc. The documentation should also identify the type of training or certification the test administrator needs to acquire and how frequently the administrator should be recertified in order to stay current with changes in the test or in the test administration.

Standard 5.6

When applicants are provided with their test scores, the communication should provide enough information that they can interpret their scores relative to the purpose for which the test was administered.

6. Testing Persons with Disabilities

Canadian law guarantees equal opportunities for all members of society. Therefore, in an appointment process, all applicants being assessed, including those with disabilities, must be provided an equitable opportunity to demonstrate their qualifications. Nonetheless, equitable assessment does not necessarily require the use of the same assessment procedure for all applicants. For example, in some circumstances, such as it sometimes occurs in the assessment of applicants with disabilities, equitable assessment will require changes or modifications to the usual testing procedure or to the test format or content. These changes or modifications are commonly called assessment accommodations.

Assessment accommodations enable persons with disabilities to participate in assessment on the same level playing field as their peers. Assessment accommodations provide an opportunity for persons with disabilities to demonstrate their qualifications without being unfairly limited or restricted due to the effects of a disability. Assessment accommodations are not to substitute for qualifications that a person does not have. They should not be excessive and should alter the standard administration of an assessment instrument to the least extent possible. In addition, to preserve the validity of results, assessment accommodations must not modify the nature or level of the qualification being assessed. This is essential to the application of merit in the appointment process.

When changes or modifications made to the testing process or to the content or format of the assessment instrument itself modify the nature or level of the qualification being assessed, these changes or modifications should not be used as testing accommodations. This awareness that some assessment accommodations can modify the nature or level of the qualification being assessed is critical when determining appropriate assessment accommodations. In fact, depending on what is being assessed, testing accommodations which can be appropriate in some circumstances may result in a modification of what is being assessed in other circumstances. An example of this would be the following:

> Although the use of a dictionary may be appropriate for a test assessing the qualification "knowledge of the organization mandate and its business", its use could be inappropriate for a test assessing the qualification "communicate in writing".

Therefore, when making a decision regarding appropriate testing accommodations, it is important to know the purpose of the test and what it assesses. Assessment accommodations that modify what is being assessed yield scores that do not provide the assessment board with valid information upon which to guide their appointment decision.

The document "Guide for Assessing Persons with Disabilities" produced by the Public Service Commission of Canada, provides a framework of principles and recommended procedures which those in charge of designing testing accommodations can use in handling concrete cases.

Standard 6.1

When assessing persons with disabilities, the need to assess the qualification required for the job cannot be waived. Rather, testing should proceed using assessment accommodations. Accommodations used in a particular assessment situation must be designed in a way that preserves the nature and level of the qualification being assessed, and maintains the standardized nature of the test.

Standard 6.2

Those in charge of designing testing accommodations when assessing a person with a disability must have sufficient knowledge of the test to be used and of what the test assesses, in order to make their decision. They must also know how the performance on the test is may be affected by the functional limitations resulting from the person's disability.

Standard 6.3

Assessment accommodations must be determined on a case-by-case basis given the wide variation in the nature and degree of functional limitations in relation to a disability and the different types of tests that can be used.

Standard 6.4

Assessment accommodations must not provide a disadvantage or an unfair advantage for persons with disabilities over other persons to be assessed using the same tests.

Standard 6.5

A record of the assessment accommodations provided and their justification must be kept.

7. Test Administration, Scoring and Security

For test use to be equitable and test scores accurate, it is important that they be administered under the conditions established by the test developer and with due sensitivity to the interaction between the test administrator and test taker. Administering tests under standardized conditions minimizes the influence of irrelevant differences in testing procedures on test scores. The meaningfulness of test score interpretation also depends on the accurate scoring of the test and the accurate reporting of the results.

Test results represent important confidential information about a person. Test users must put in place a system to maintain the security of test results and to limit access to these results to authorized persons.

Standard 7.1

Applicants must be assessed in English or French, or both, according to their choice, except for the purpose of determining language proficiency or assessing language-related skills.

Standard 7.2

Test takers should be treated fairly and impartially during the assessment process. Test administrators should treat all test takers courteously and with respect, and should carefully follow the standardized procedures for test administration and scoring specified by the test developer.

Standard 7.3

Test takers should have access in advance to information about the nature of a test and how it will be used. Access should be provided equally to all persons who will be taking the test.

Standard 7.4

The testing environment should furnish reasonable comfort with minimal distraction. Instructions to test takers should clearly indicate how to make responses. Testing materials should be readable and understandable.

Standard 7.5

Efforts should be made to monitor the testing situation so that opportunities for cheating are minimized.

Standard 7.6

Test users have the responsibility of protecting the security of test materials at all times.

Standard 7.7

Procedures need to be developed to ensure the accuracy of test scoring. Frequency of scoring errors should be monitored. Adherence to established scoring criteria should be checked regularly.

Standard 7.8

The confidentiality of individual test results should be ensured. Test users who maintain test scores on individuals in data files or in an individual's records should develop a clear set of policy guidelines on the duration of retention of an individual's record, and on the availability, and use over time, of such data. Test results should be stored in a manner which permits responses to legitimate requests for scores and the correction of erroneous data.

PART B: LEGAL ISSUES CONCERNING TEST USE

The purpose of this section is to provide information on the legal basis for using standardized tests in staffing actions in the Public Service. Reference is made to the legislation governing staffing in the Public Service, namely, the *Public Service Employment Act* (PSEA) and the *Public Service Employment Regulations* (PSER), and other legislations which has an impact on test use for appointment.

This section describes the legal role of the Public Service Commission to appoint qualified people to the Public Service, and to select assessment instruments, including tests. The rights of individuals to be tested in either or both official languages are discussed. Also outlined are the legal sanctions against fraudulent practices.

The Commission can establish policies concerning appointments. The policies that have a direct impact on the use of tests will be described.

According to the PSEA, the employer may establish qualification standards, in relation to education, knowledge, experience, occupational certification, language or other qualifications. These will also be discussed.

The right of the Public Service Commission to maintain the confidentiality of test materials and the right of individuals to gain access to tests or testing materials is described with reference to the PSER, the *Access to Information Act* and *the Privacy Act*.

1. The Public Service Commission's Authority to Staff Positions in the Public Service

The *Public Service Employment Act* (PSEA) is the legal basis for the authority of the Public Service Commission to provide for the appointment of qualified persons and to delegate its powers, functions and duties in this regard.

11. "The mandate of the Commission is

    (a) to appoint, or provide for the appointment of, persons to or from within the public service in accordance with this Act;"

15. (1) "The Commission may authorize a deputy head to exercise or perform, in relation to his or her organization, in the manner and subject to any terms and conditions that the Commission directs, any of the powers and functions of the Commission under this Act, …. "

Section 30 defines merit, the basis on which appointments are made:

30. (1) "Appointments by the Commission to or from within the public service shall be made on the basis of merit and must be free from political influence.

(2) An appointment is made on the basis of merit when:

(a) the Commission is satisfied that the person to be appointed meets the essential qualifications for the work to be performed, as established by the deputy head, including official language proficiency; and

(b) the Commission has regard to

(i) any additional qualifications that the deputy head may consider to be an asset for the work to be performed, or for the organization, currently or in the future,

(ii) any current or future operational requirements of the organization that may be identified by the deputy head, and

(iii) any current or future needs of the organization that may be identified by the deputy head."

2. The Public Service Commission's Authority to Select a Test

The authority of the Public Service Commission and, through delegation, of management to select and use tests in making appointments is found in Section 36 of the Act which reads as follows:

36. "In making an appointment, the Commission may use any assessment method, such as a review of past performance and accomplishments, interviews and examinations, that it considers appropriate to determine whether a person meets the qualifications referred to in paragraph 30(2)(a) and subparagraph 30(2)(b)(i)."

3. Testing and the Official Languages

Subsections 37(1) and 37(2) of the *Public Service Employment Act* specifies the language in which tests should be administered:

37. (1) "An examination or interview, when conducted for the purpose of assessing qualifications referred to in paragraph 30(2)(a) and subparagraph 30(2)(b)(i), other than language proficiency, shall be conducted in English or French or both at the option of the candidate.

(2) An examination or interview, when conducted for the purpose of assessing the qualifications of the candidate in the knowledge and use of English or French or both, or of a third language, shall be conducted in that language or those languages."

In short, this subsection of the Act grants applicants who are to be administered a selection test the right to take it in English or in French or in both languages, as they wish, except in the case of a language proficiency test or a technical skill test where the content is intrinsically related to the knowledge and use of language. In the latter case the applicant

has no choice: he or she must take the test in the language (or languages) in which that proficiency or that technical language-related skill is required in order to perform the duties of the position.

4. Fraudulent Behaviour During the Testing Process

Sections 69 and 133 of the PSEA prohibit the use of PSC tests and testing materials for fraudulent purposes. Applicants attempting to increase their scores on tests by copying or otherwise obtaining protected test materials, by referring to unauthorized materials while writing the examination, by copying from another person, by impersonating someone else, or by engaging in any other form of cheating constitute fraudulent behaviours.

69. "If it has reason to believe that fraud may have occurred in an appointment process, the Commission may investigate the appointment process and, if it is satisfied that fraud has occurred, the Commission may

    (a) revoke the appointment or not make the appointment, as the case may be; and

    (b) take any corrective action that it considers appropriate."

133. "Every person who commits fraud in any appointment process is guilty of an offence punishable on summary conviction."

5. The Authority of the Public Service Commission to Establish Policies

The Commission can establish policies concerning appointments according to Subsection 29(3) of the PSEA.

29. (3) "The Commission may establish policies respecting the manner of making and revoking appointments and taking corrective action."

The following is the policy statement concerning assessment of applicants:

- The assessment is designed and implemented without bias, political influence or personal favouritism and does not create systemic barriers.

- The assessment processes and methods effectively assess the essential qualifications and other merit criteria identified and are administered fairly.

- The identification of persons who meet the operational requirements and organizational needs is carried out objectively.

Requirements:

In addition to being accountable for respecting the policy statement, deputy heads must:

- inform the persons to be assessed, in a timely manner, of the assessment methods to be used, their right to accommodation and how to exercise that right;
- and ensure that those responsible for assessment:
  - have the necessary competencies to ensure a fair and complete assessment of the person's qualifications;
  - have the language proficiency required to permit effective communication with the person being assessed in the official language or languages chosen by that person in order to assess his or her qualifications fairly;
  - are not in conflict of interest and are able to carry out their roles, responsibilities and duties in a fair and just manner;
  - obtain the PSC's approval before using tests of personality, intelligence, aptitude, or tests of like nature;
  - adhere to the guidelines set forth in the document entitled "Testing in the Public Service of Canada", published by the PSC, when developing and using standardized tests;
  - use assessment tools that do not create systemic barriers to employment;
  - use the PSC's Second Language Evaluation test or another instrument approved by the PSC to assess official language skills on a "meets/does-not-meet" basis. For appointments of students or casual workers, those responsible for the assessment are permitted to conduct the assessment if they have the language skills required to do so;
  - conduct their own assessment of expert or specialized official language proficiency qualifications;
  - assess qualifications for appointment to or within the EX group:
    - with a structured interview and a structured reference check, and;
    - with any additional assessment tools necessary to provide clear evidence for appointment decisions (such as SELEX)
    - obtain approval from the PSC for exceptions to the EX assessment requirements on a case-by-case basis
  - establish EX assessment boards comprised of members at, or equivalent to, a level above the position being staffed.

Policy requirements concerning Employment Equity in the Appointment Process will be presented in section 8 of this document as it relates to the *Canadian Human Rights Act* and *the Employment Equity Act*.

6. The Employer's Responsibility to Establish Qualification Standards

The PSEA gives the Employer the responsibility to establish qualification standards.

31. (1) "The employer may establish qualification standards, in relation to education, knowledge, experience, occupational certification, language or other qualifications, that the employer considers necessary or desirable having regard to the nature of the work to be performed and the present and future needs of the public service.

(2) The qualifications referred to in paragraph 30(2)(a) and subparagraph 30(2)(b)(i) must meet or exceed any applicable qualification standards established by the employer under subsection (1)."

The Public Service Human Resources Management Agency of Canada (PSHRMAC) develops and keeps up to date these standards on behalf of the employer. They include the Occupational *Group Qualification Standards* and the *Qualification Standards in Relation to Official Languages*

The *Occupational Group Qualification Standards* outline the minimum requirements (in terms of education, occupational certification, etc.) to be used when appointing to positions in the public service. These standards sometimes include mandatory tests for some professional groups. While the employer is responsible for establishing and maintaining the qualification standards, the Public Service Commission is responsible for the assessment of qualifications. In this capacity, the PSC prescribes or approves tests that are referred to in the Occupational Group Qualification Standards.

The types of qualifications that could be specified on the Statement of Merit Criteria are defined in these standards. The following types of qualifications are recognized:

Education
Knowledge
Experience
Occupational Certification
Official Language Proficiency
Abilities/Skills
Aptitudes
Personal Suitability

Tests as Alternatives to Education

The Qualification Standards specify the approved alternatives to formal education. Some PSC tests can be used as alternatives to education. They allow individuals who have not achieved a specified educational level an opportunity to demonstrate that they meet the educational standards prescribed in qualification standards. These alternatives are valid and relevant only in terms of the Qualification Standards of the Public Service. In no other way is the successful completion of a Public Service Commission test that is used as an

alternative to education can be recognized as a diploma, certificate, or degree obtained in good and due form.

Tests used as alternatives to formal education are available for the following educational levels:

- successful completion of two years of secondary school;
- a secondary school diploma; and
- university graduation (Groups of the Scientific and Professional Category excepted).

The *Qualification Standards in Relation to Official Languages* outline the levels of official language proficiency qualifications required for bilingual positions in the core public administration and prescribes the use of PSC approved tests to this effect.

7. Access to Tests and Testing Materials

The *Access to Information Act* explicitly restricts the individuals' opportunity to examine tests and related materials in order to ensure the valid use of the test. Section 22 of the Act reads:

22. "The head of a government institution may refuse to disclose any record requested under this Act that contains information relating to testing or auditing procedures or techniques or details of specific tests to be given or audits to be conducted if such disclosure would prejudice the use or results of particular tests or audits."

The PSC gave explicit protection to standardized tests that could be disclosed in the course of an investigation.

The *Public Service Employment Regulation* (PSER) stipulates that:

20. (1) "The Commission shall not disclose a standardized test, or information concerning a standardized test, owned by an organization or the Commission or that is commercially available, if obtained in the course of an investigation under the Act, unless it can be disclosed, with or without conditions set by the Commission, in a manner that will not affect the validity or continued use of the standardized test or will not affect the results of such a test by giving an unfair advantage to any person."

The PSER explicitly defines a standardized test.

20. (2) "For the purpose of subsection (1), a standardized test is a systematic procedure for sampling an individual's behaviour in order to evaluate job-relevant competencies. The procedure is systematic in five areas: development, content, administration, scoring and communication of results. The content of the test is equivalent for all test-takers. The test is administered according to standard instructions and procedures and is scored according to a set protocol."

The Public Service Staffing Tribunal, through regulations, gave similar protection to standardized tests that could be disclosed in the context of a complaint. This is the relevant section of *the Public Service Staffing Tribunal Regulations*:

17. (1) "Despite section 16, the complainant or the deputy head or the Commission may refuse to provide information referred to in that section if providing that information might:
(a) threaten national security;
(b) threaten any person's safety; or
(c) affect the validity or continued use of a standardized test or parts of the test or affect the results of a standardized test by giving an unfair advantage to any individual."

Thus, test booklets, scoring keys, item banks and any other documents or materials which might compromise the security of the test are disclosed only under conditions that would not affect the validity or continued use of a test or would not give an unfair advantage to any individual.

As specified in Section 12 of the *Privacy Act*, individuals may have access to personal records, subject to various conditions and regulations. Materials which contain personal information gathered during a testing session and deemed not to undermine the security of any PSC or departmental test or testing procedure may be made available as described in Subsection 17(1) of the *Privacy Act* which reads:

17. (1) "Subject to any regulations made under paragraph 77(1)(o), where an individual is to be given access to personal information requested under subsection 12(1), the government institution shall
(a) permit the individual to examine the information in accordance with the regulations; or
(b) provide the individual with a copy thereof."

As mentioned above, this doesn't imply that test developers are required to divulge tests or testing materials deemed to undermine test security. However, under certain circumstances, applicants may be permitted to view certain documents, such as their answer sheets. Test developers are not required to make copies of such documents available to the applicant. Indeed, copies of test materials which would compromise the confidentiality of the assessment instrument will not be made available to safeguard the security of the test and to ensure its valid application.

8. Two Acts That Have an Important Impact on Test Use in the Public Service

The Canadian Human Rights Act (1998)

The purpose of this Act is : "…to extend the laws in Canada to give effect, … to the principle that all individuals should have an opportunity equal with other individuals to make for themselves the lives that they are able and wish to have and to have their needs accommodated, consistent with their duties and obligations as members of society, without being hindered in or prevented from doing so by discriminatory practices based on race, national or ethnic origin, colour, religion, age, sex, sexual orientation, marital status, family status, disability or conviction for an offence for which a pardon has been granted." (Section 2)

This act emphasizes the obligation to provide accommodations, including the personnel selection process in which tests are being used.

The Employment Equity Act (1995)

The purpose of the *Employment Equity Act* (EEA) is: "...to achieve equality in the workplace so that no person shall be denied employment opportunities or benefits for reasons unrelated to ability and, in the fulfillment of that goal, to correct the conditions of disadvantage in employment experienced by women, aboriginal peoples, persons with disabilities and members of visible minorities by giving effect to the principle that employment equity means more than treating persons in the same way but also requires special measures and the accommodation of differences**.** (Section 2)

The EEA requires employers to identify and remove barriers to employment of persons in the four designated groups, and to institute positive policies and practices and make reasonable accommodation*s* to ensure that persons in the four designated groups achieve representation in the employer's workforce proportionate to their labour market availability. (Section 5)

In line with these Acts, there are specific requirements in the PSC Appointment Framework related to Employment Equity in the Appointment Process:

In addition to being accountable for respecting the policy statement, deputy heads must:

- accommodate the needs of persons through all stages of the appointment process to address, up to the point of undue hardship, disadvantages arising from prohibited grounds of discrimination;
- use assessment tools and processes that are designed and implemented without bias and do not create systemic barriers.