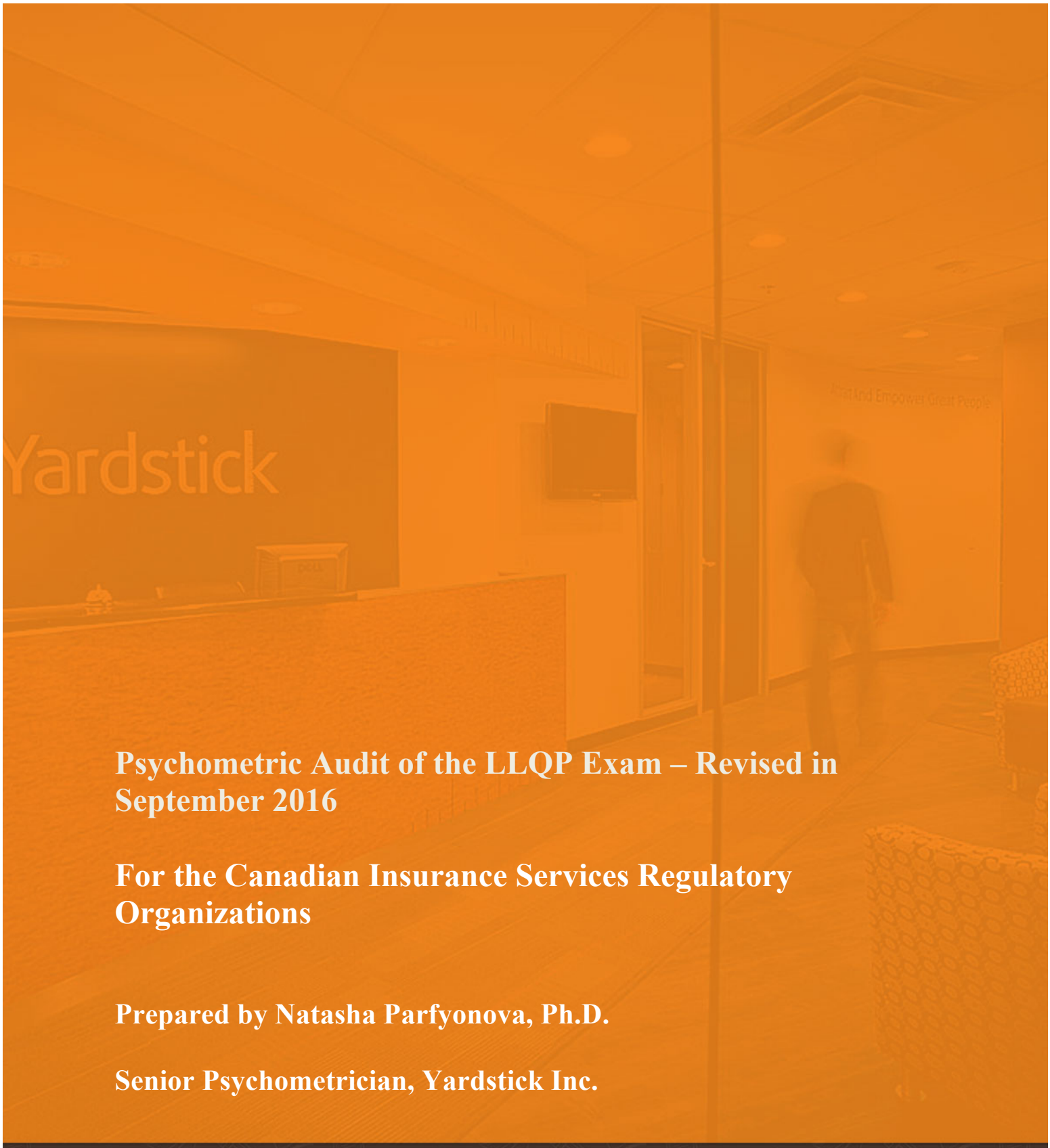

Psychometric Audit of the LLQP Exam – Revised in September 2016

For the Canadian Insurance Services Regulatory Organizations

Prepared by Natasha Parfyonova, Ph.D.

Senior Psychometrician, Yardstick Inc.

Yardstick
Testing and Training Experts

Table of Contents

Executive Summary

The Canadian Insurance Services Regulatory Organizations (CISRO) contracted a testing and training firm, Yardstick, to complete a psychometric audit of the Life License Qualification Program (LLQP) exam. Based on the information gathered during the audit, the current processes used to develop, administer, and score the LLQP exam conform to the majority of testing standards. CISRO used psychometrically sound procedures to identify exam content and specifications and create exam items. Considerable effort was involved in item writing, validation, and translation.

The psychometric review of the English and French exam items established that they were written in accordance with the general principles of item writing. The direct comparison of difficulty levels of the current and old LLQP exams was not possible because the current exam contains new items.

The statistical analyses of exam and item performance were conducted for two versions of the English LLQP exam forms administered in the spring of 2016. The statistical analyses revealed that exam forms have levels of reliability that are expected of short exams, and that the majority of exam items have adequate power to differentiate between low- and high-scoring examinees. The results of statistical analyses of alternate exam forms are consistent across the two exam versions.

To enhance the LLQP exam development and administration program, Yardstick recommends a number of improvements, discussed in detail in this report. There are, however, two key areas that, without improvement, can undermine the validity of exam score interpretation.

- One of these areas is exam administration, scoring, and reporting. Exam scores are more likely to have the same meaning across examinees when an exam is administered under the same examination conditions. This suggests the standardization of exam administration, scoring, and reporting policies and procedures for the LLQP exam in all exam locations, to promote equal treatment of examinees and consistent score interpretation. Without standardized policies and procedures in place, it is difficult to ensure accurate measurement of examinees' competencies. Yardstick recommends that all jurisdictions follow the same policies and procedures for exam administration, use the same scoring protocol, and report exam results to examinees in a consistent manner.

- Another area for improvement is standard setting. Standard setting refers to the process by which an exam passmark is set. The current passmark was set more than a decade ago, and due to the method of standard setting used at that time, it is not tied to any performance standard. Consequently, the passmark is not reflective of the current performance requirements for entry-level life insurance agents and does not take into consideration the difficulty of exam items. It is recommended that CISRO conduct a new standard setting study to obtain a new exam passmark that is reflective of industry expectations for the knowledge and skills of life insurance agents at job entry.

The original report was released to CISRO stakeholders for review in June 2016. In September 2016, the report was revised to include additional information in response to feedback and questions from multiple stakeholders.

Introduction

In January 2016, Yardstick was contracted to conduct a psychometric audit of the LLQP exam for CISRO. The objective of the audit was to evaluate the LLQP exam against testing standards to determine how effectively the exam achieves its purpose of identifying competent entry-level life insurance agents. As a result of this audit, CISRO will be able to determine how its exam development and administration processes compare to testing standards, and where any improvements could be made.

As part of the exam audit, Yardstick reviewed evidence of exam reliability and fairness, as well as evidence supporting the validity of conclusions drawn from exam scores. The standards for evaluation of the LLQP Exam were derived primarily from the benchmark publication in the testing industry, titled the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 2014). The *Standards for the Accreditation of Certification Programs* issued by the National Commission for Certifying Agencies (NCCA, 2014) were also used as appropriate. In this report, the standards from the first document are referred to as the *Standards,* while the standards from the second document are referred to as the *NCCA Standards*.

The exam audit was divided into six stages of exam development. Each stage contributes to the quality of the LLQP Exam and has a bearing on the accuracy of conclusions drawn from exam scores. The six stages of exam development are briefly described below and more fully elaborated on in the subsequent sections of this report.

STAGE I: Exam Purpose, Content, and Specifications

The first stage of exam development deals with defining exam purpose, identifying the appropriate exam content and deciding on the appropriate way to assess it. An exam should target competencies that are important for the competent practice of professionals. Moreover, these competencies should be assessed in a way that allows for meaningful conclusions to be drawn from exam scores. The first stage of exam audit focuses on the processes involved in the development and validation of the *Life Insurance Agent Competency Profile* and the LLQP Exam specifications.

STAGE II: Item Development

The second stage of exam development involves the creation of a robust item bank of exam items that follow best practice in item development and meet exam specifications. This audit evaluated CISRO multiple-choice items against the principles of item writing and examined how effectively CISRO's item development processes enabled the organization to fulfill the exam specifications and the overall needs of the exam program.

STAGE III: Exam Assembly

Exam assembly refers to processes underlying the construction of comparable exam forms from items in the item bank. The goal of exam assembly is to ensure that exam forms are in alignment with exam specifications and comparable in terms of their difficulty. The

specific processes in place for exam assembly are evaluated, as well as the evidence for adherence of alternate exam forms to the exam specifications.

STAGE IV: Exam Administration, Scoring, and Reporting

Exam administration refers to the processes through which exam forms are administered to examinees, while exam scoring and reporting includes processes for exam scoring and communication of exam results to examinees. The audit examines the extent to which exam administration, scoring, and reporting processes are standardized, complete, and secure.

STAGE V: Exam and Item Analysis

Exam and item analysis investigates psychometric properties of an exam and exam items through statistical means to determine how effectively the exam differentiates between examinees based on their competence level. A good exam will measure examinees' competence consistently. It will also allow the differentiation between competent and not – yet–competent examinees. This section includes the review of statistical criteria used by CISRO to evaluate the quality of LLQP questions. Also, it includes the results of exam and item analyses of the LLQP modular exams completed by Yardstick.

STAGE VI: Standard Setting

Standard setting is the process by which an exam passmark is established and validated. For exam scores to be meaningful, the passmark needs to be established through a scientific process that considers item content, item difficulty, and the expected performance of a minimally competent examinee. The exam audit evaluates the standard setting process used to set a passmark on the LLQP Exam and the documentation of the outcomes.

The information required for the LLQP Exam audit was obtained from the written documentation provided by CISRO, as well as from interviews with the CISRO exam development team from the Autorité des Marchés Financiers (AMF). The findings and recommendations presented in this audit are based on the availability of information provided by CISRO and its exam development team. The following documents were used in this audit:

- Autorité des Marchés Financiers (2010, May). *The Autorité des marchés financiers' process for developing examinations: Applying the best docimological practices.*
- Autorité des Marchés Financiers (2012, Spring). *Occupational analysis report.*
- CISRO/ORCA (2013, June). *Competency profile: Life Insurance Agent.*
- CISRO/OCRA (2013, June). *Survey results overview. Competency profile: Life Insurance Agent.*
- CISRO/OCRA (2013, November). *Report on transitional measures: Implementation of new LLQP exam.*
- CISRO/OCRA (2014, March). *Report on LLQP curriculum survey results.*
- CISRO/ORCA (2014, June). *Life License Qualification Program: Frequently asked questions.*
- CISRO/OCRA (2014, August). *Control of exam validity: Implementation of harmonized LLQP.*
- CISRO/OCRA (2014, September). *Life Licence Qualification Program (LLQP): Guidelines for the implementation of the LLQP.*
- CISRO/OCRA (2014, December). *Guidelines for drafting and review: Harmonized LLQP exam*

questions.
- CISRO/ORCA (2014). *Measurement and evaluation roadmap - Fall 2014: Creation and implementation of new formats.*
- CISRO/OCRA (2015, May) *Curriculum: Life Licence Qualification Program (LLQP).*
- CISRO/OCRA (2015, December). *Practical exam administration guidelines: Life Licence Qualification Program (LLQP).*
- CISRO/ORCA (n.d.). *Measurement and evaluation analyst procedures.*
- CISRO/ORCA (n.d). *Harmonized qualification standards. Subject matter needs and requirements*.
- Dickson, M., & Hultgren, D. (2001). *Description of issues affecting the calculation of pass scores for the sub-tests of the LLQP certification exam's initial administrations.*
- *Grille de Construction des Examens* (n.d.).
- *LLQP exam monitoring: Transitional and ongoing process* (2016, May).

The original psychometric audit report was released to CISRO stakeholders for review. In response to feedback and questions from stakeholders, the report was amended to include Appendix A that provides additional information on the LLQP exam, including the results of comparison of pass rates and difficulty levels for the old and new exams.

Stage I –Exam Purpose, Content, and Specifications

During Stage I, Yardstick examined the purpose of the LLQP exam and the processes involved in the development of the *Life Insurance Agent Competency Profile* and exam specifications. The goal of this review was to identify how well the content of the exam is aligned with exam purpose and specifications.

Clearly stating the purpose(s) of the exam is the first and the most fundamental step in exam development and evaluation of results. A good exam has a clearly stated purpose that is aligned with the intended interpretation of exam scores.

Exam Purpose and Intended Uses of Scores

Standard 1.1 states that an exam developer should clearly identify the construct that an exam measures, the exam population that the exam applies to, and the purpose for which exam scores will be used. *Standard 1.2* suggests that an exam developer should provide an explanation of why exam scores can be used for the intended purpose.

> ***Standard 1.1*** The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

> ***Standard 1.2*** A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.

The LLQP exam is a licensing qualification standard that applies to all individuals in Canada who want to become licensed as life insurance agents. Only individuals who completed a relevant training program can challenge the LLQP exam. Following successful completion of the exam, examinees obtain a certificate and are eligible to apply for a life insurance license with the provincial regulator.

As stated in the document titled *Life Licence Qualification Program (LLQP): Frequently Asked Questions*, the LLQP Exam was developed by CISRO through a collaborative process with Canada's insurance regulators and its purpose is to "protect consumers by helping ensure agents are financially literate about life insurance products" (p. 1). According to the document titled *Curriculum: Life Licence Qualification Program (LLQP)*, the purpose of the LLQP Exam is to assess examinees' competence to practice ethically in accordance with consumer rights" (p. 1).

The construct that the LLQP exam is intended to assess is described in the *Life Insurance Agent Competency Profile* and the LLQP exam specifications that are known in the industry under the name *Curriculum*. The LLQP exam consists of four modular exams that assess examinees' competencies in

the following areas: 1) Ethics, 2) Life Insurance, 3) A&S Insurance, and 4) Segregated Funds. The LLQP Exam assesses competencies required of life insurance agents upon career entry.

According to the exam development manual titled *The Autorité des marchés financiers' process for developing examinations: Applying the best docimological practices,* the LLQP exam is an open-book exam. In other words, examinees can look up answers in the reference material provided. This information further defines exam purpose and the construct that is being assessed by the exam.

Competency Profile

Defining the content of the future exam is one of the fundamental steps of exam development. The validity of inferences made on the basis of exam scores depends on the extent to which exam content reflects the construct the exam is supposed to assess. The better the alignment between exam content and the construct of interest, the more likely it is that exam scores can be interpreted in a meaningful way.

The LLQP exam, for example, is supposed to assess critical competencies of entry-level life insurance agents that help them practice in a competent and ethical manner. The validity of interpretation of exam scores depends on the availability of evidence that suggests that exam questions, indeed, tap into the competencies required of entry-level life insurance agents. This evidence relies on the judgment of subject matter experts obtained in the process of exam development and validation.
According to *Standard 1.11*, an exam developer is required to describe and document the processes underlying the development of competencies for the intended examinee population. Additionally, the criteria for competency selection, such as importance, frequency or criticality, should be specified.

> ***Standard 1.11*** When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

The information on competency development was obtained from the following documents: 1) *Life Insurance Agent Competency Profile*; 2) *Survey Results Overview*; and 3) exam development manual. The *Life Insurance Agent Competency Profile* was developed in 2012, using a two-tiered process that involved extensive consultations with subject matter experts in five provinces; Alberta, British Columbia, Ontario, New Brunswick, and Quebec. There were three occupational analysis workshops with subject matter experts in the following occupations: representatives in insurance of persons, accident and sickness insurance representatives, and representatives in group insurance of persons. The competency validation process consisted of an administration of a competency validation survey.

Standard 1.8 calls for a thorough description of samples used for validation purposes, including participants' experience, qualifications and socio-demographic characteristics that are relevant to the purpose of the exam. *Standard 1.9* reiterates the importance of specifying procedures for selection and training of subject matter experts, which includes instructions given to them in data collection tasks and the processes they used to reach decisions. *Standard 7.5* specifies a requirement to document it all.

Standard 1.8 The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant socio-demographic and developmental characteristics.

Standard 1.9 When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Standard 7.5 Test documents should record the relevant characteristics of the individuals or groups of individuals who participated in data collection efforts associated with test development or validation (e.g., demographic information, job status, grade level); the nature of the data that were contributed (e.g., predictor data, criterion data); the nature of judgments made by subject matter experts (e.g., content validation linkages); the instructions that were provided to participants in data collection efforts for their specific tasks; and the conditions under which the test data were collected in the validity study.

According to the exam development manual, the purpose of the occupational analysis workshops was to: 1) uncover tasks performed by life insurance agents; 2) derive knowledge, skills, and abilities necessary to carry out those tasks; and 3) determine qualification requirements for life insurance agents. The workshops followed the standard occupational analysis methodology approved by the federal and provincial governments. A detailed description of this methodology can be found in the document titled *Occupational Analysis Report.* That document provides information on the selection criteria for workshop participants, their qualifications, the tasks they completed in the workshop, and the outcomes of discussions. It is worth noting that socio-demographic characteristics and experience of the workshop participants were not documented.

As a result of occupational analysis workshops, an initial draft of the national *Life Insurance Agent Competency Profile* was created. The document includes a list of "tasks and operations that a life insurance agent can accomplish upon career entry" (i.e., within three years of starting on the job). (p. 5) The *Life Insurance Agent Competency Profile* consists of three major sections: *Areas, Competencies,* and *Competency Components*.

After the workshops, an online survey was conducted to validate the content of the *Life Insurance Agent Competency Profile* with a broader sample of subject matter experts. The survey was available in both official languages and was open to various roles in the industry, including but not limited to agents, brokers, provincial regulators, course providers, business managers, and human resource professionals.

A total of 751 respondents completed the survey. In line with expectations, the largest groups of respondents who completed the survey were agents (58%) and brokers (29%). It is difficult to judge if survey results generalized to all stakeholders in the country given a very uneven distribution of survey respondents across provinces. As acknowledged in the survey report, 72% of survey respondents came from Alberta even though it is not the largest province in terms of the number of life insurance agents who work there. The competency developers compared the survey results for Alberta with those for the other provinces, and concluded that they were comparable.

The survey respondents were asked to make two holistic judgements about the entire competency profile (e.g., "Are there relevant competencies or competency components that cannot be found in the competency profile?" and " Are any of the profile's competencies or competency components NOT relevant to professional practice?"). More than 90% of respondents indicated that the document was complete and 98% answered that all of the competencies were relevant. The qualitative feedback from the survey was used to make modifications to the *Life Insurance Agent Competency Profile*.

Exam Specifications

Once the competency profile for the exam is developed, the next step is to translate competencies into exam specifications. Exam specifications provide a detailed roadmap for exam construction and serve as evidence of content-related validity required to support exam score interpretations.

Standard 4.1 recommends that exam specifications include a description of the exam purpose, the intended examinee population, the construct that is being measured, as well as the intended uses of exam scores. In addition, the *NCCA Standard 15* recommends that the appropriate level of practice for examinees (e.g., entry, advanced) be specified.

> ***Standard 4.1*** Test specifications should describe the purpose(s) of the test, the definition of the construct or domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of test results for the intended purpose(s).

In order to harmonize the Canadian life insurance qualification process, CISRO created a new set of exam specifications that are also known as the LLQP Exam Curriculum. The exam specifications represent a number of evaluation tables that contain competencies and competency components that must be evaluated on the LLQP Exam. Each competency is measured with a modular exam, and all modular exams are equally weighted in the assessment process. In order to pass the LLQP Exam, an examinee must score at or above the passmark on all four modular exams.

Each modular exam evaluates two or four competency components, which are weighted according to their relative importance for consumer protection, as well as the complexity of the underlying concepts. Each competency component is further broken down into sub-components, the measurement of which is limited to a list of topics referred to as "related contents." In addition, the exam specifications also specify the administration time and length of each modular exam.

The process underlying the development of exam specifications is described in the exam development manual. The exam specifications were developed by a group of measurement and evaluation specialists,

trainers, and practitioners using the content of the *Life Insurance Competency Profile*. The socio-demographic information and qualifications of group members were not documented in the exam development manual. However, Yardstick knows that CISRO collects socio-demographic information on all volunteers who contribute to exam development and validation efforts.

The group selected competency components to be assessed on the exam and "determined relative importance of each component in the make-up of the skill" (p. 5). The competency component selection was guided by the following criteria: 1) relevance to the CISRO's mission to protect consumers; 2) appropriateness for entry-level practitioners; and 3) suitability for measurement with multiple-choice questions. The process used to identify relative importance of competency components was not documented. It appears that the competency weights were set using the professional judgment of subject matter experts.

Standard 4.2 states that the information on exam content, proposed exam length, item formats, time limit, as well as directions to examinees and scoring and reporting procedures should be provided.

> ***Standard 4.2*** In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. Test specifications should also specify the amount of time allowed for testing; directions for the test takers; procedures to be used for test administration, including permissible variations; any materials to be used; and scoring and reporting procedures. Specifications for computer-based tests should include a description of any hardware and software requirements.

The LLQP exam specifications include such important information about the exam as exam weights by competency component, the amount of time allowed for each exam, and exam length. Yet, there is no mention of other critical information, such as item type (e.g., multiple-choice items with four answer options), the availability of the exam in both official languages, and scoring and reporting procedures.

According to the *NCCA Standard 15,* an exam developer should document the following information in exam specifications:

- Exam purpose;
- Description of the examinee population;
- Description of the construct and item types to be used;
- The weighted content outline
- Criteria for exam assembly
- Exam administration requirements (e.g., a computer-based exam, the availability of reference material, the use of calculators)
- General description of the plan for scoring and equating the exam and for running the psychometric analysis.

The next stage in developing the LLQP exam specifications included a stakeholder survey. As required by *Standard 4.6,* CISRO and provincial regulators invited stakeholders from the industry to evaluate the

appropriateness of exam specifications through the online survey. The evaluation process and its results are described in the document titled *Report on LLQP Curriculum Survey Results.*

Standard 4.6 specifies that purpose of the exam specifications review, the process by which it was conducted, the results of the review, and the qualifications, experiences, and demographic characteristics of reviewers should be documented.

> ***Standard 4.6*** When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.

The requirements of *Standard 4.6* were met. The exam specifications survey report describes the process by which the link to the survey was disseminated and provides insight into the characteristics of the sample of respondents. The survey was completed by 386 respondents, which represented agents (59%), brokers (23%), and other related professions, such as provincial regulators, sales managers, etc. (18%). The only demographic information available on the respondents was their province of residence. Similar to the competency validation survey, the respondents from Alberta were overrepresented in the sample (69%) and the respondents from Ontario were underrepresented (13%). The survey developers said that they compared the overall survey results to those provided by respondents other than Alberta and found no differences. The survey report contains survey questions and describes survey results in detail.

Summary and Recommendations

- ***Purpose and Intended Uses of Exam Scores.*** The purpose of the LLQP Exam, the intended examinee population, the construct that is being assessed, and the intended use of exam scores is currently provided in different technical documents. To comply with *Standards 1.1 and 1.2,* it is important that CISRO clearly articulates this information in the exam development manual and the documents that are available to the public (e.g., an examinee handbook).

- ***Competency Profile.*** Psychometrically sound procedures were used to develop the exam content domain; i.e., Life Insurance Agent Competency Profile. The procedures involved occupational analysis workshops and the online validation survey with a broad sample of subject matter experts in the appropriate professional roles. The requirements of *Standards 1.8. 1.9, 1.11,* and *7.5* for documenting the process underlying the competency profile development were met. CISRO has robust documentation on the process used to conduct occupational analysis workshops and the competency validation survey. The *Standards* call for the documentation of socio-demographic characteristics of samples used in validity studies. It should be noted that the *Standards* were published in the United States and some standards may not apply to the Canadian legal context. It is sufficient to document geographic representation of subject matter experts, their qualifications, work experience, and areas of expertise.

- ***Exam Specifications.*** CISRO used a sound psychometric procedure to develop and validate detailed exam specifications. The competencies to be assessed on the exam were derived from the *Life Insurance Agent Competency Profile,* and all key decisions on the content and structure of the LLQP exam were made in consultation with the industry. The intended examinee population and the construct that is being measured by the exam are described in great detail in the exam specifications. To fully comply with *Standard 4.1* and the *NCCA Standard 15*, CISRO should add to the exam specifications the information on item type, criteria for exam assembly, exam administration requirements (e.g., a computer-based exam, the availability of reference material, the use of calculators), and a general description of the plan for scoring and equating the exam and for running the psychometric analysis.

Stage II –Item Development

In Stage II, Yardstick evaluated the processes and procedures used to develop and review new items for the LLQP exam. Specifically, Yardstick investigated the extent to which item development processes and procedures meet standards for item development and review.

The quality of exam items determines the quality of an exam. Once exam specifications are established, items that conform to these exam specifications need to be developed.

Selection and Training of Item Writers and Reviewers

The importance of selection and rigorous training of item writers and reviewers cannot be overstated. According to *Standard 4.7*, it is important to document how item writers and reviewers are selected, what training they receive, and how they go about creating and reviewing questions. This information gives credibility to the item development process and serves as validity evidence for exam scores.

> ***Standard 4.7*** The procedures used to develop, review, and tryout items and to select items from the item pool should be documented.

The requirements for item writers and reviewers are described in detail in the document titled *Harmonized Qualification Standards: Subject Matter Needs and Requirements.* The item writers and reviewers must have a valid life insurance license obtained in Canada, sound product knowledge, sales or training responsibilities at work, interest in licensing, and excellent communication skills. In addition, they must come from different geographic regions and represent different types of organizations, including large firms, professional associations, and training institutions. Bilingual individuals are encouraged to apply. To be considered for an item writer or reviewer position, subject matter experts must submit a completed application form, a letter of intent, and their resume to CISRO.

Standard 4.8 recommends that an exam developer evaluate new items through empirical analyses and/or through the content review by subject matter experts. The training provided to the reviewers along with their qualifications and socio-demographic characteristics should be documented.

> ***Standard 4.8*** The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.

Once selected, item writers and item reviewers for the LLQP exam participate in training on item drafting and review. Based on the PowerPoint presentation provided by CISRO, the training content is aligned with the document titled *Guidelines for Drafting and Review: Harmonized LLQP Exam Questions*. The item writers and reviewers are given step-by-step instructions on how to draft or review items and provided with relevant templates.

Once an item is written, a measurement and evaluation specialist reviews it for clarity, relevance to the

exam specifications, and compliance with item-writing rules. The specialist provides feedback to the item writer and signs off on item revisions. The item is then forwarded to an item reviewer who validates the item and passes it back on to the specialist to finalize the item before it is copy edited.

In an effort to reduce bias, the review is conducted in a blind fashion where the item reviewer does not know the identity of an item writer and has to read an item that was stripped of such auxiliary information as the correct answer and the reference. The task of the item reviewer is to evaluate the item against the criteria specified in the content analysis grid. The item reviewer should verify the correct answer, item reference and link to the competency component in the exam specifications. The relevance of the item to professional practice at the entry-level, the frequency of use of content in practice, and the perceived item difficulty is also evaluated. The item reviewer is expected to suggest improvements and make changes to the item as necessary.

It can be concluded that the selection and training of subject matter experts is well documented while their characteristics are not.

Adherence to Item Writing Principles

As part of this report, a psychometric review of the LLQP exam items in the English language was conducted. As a first step, the items were reviewed against the item-writing principles published in a seminal journal article by Haladyna, Downing, and Rodriguez (2002). The issues detected in the items were recorded, and a French-speaking staff member at Yardstick identified if these issues also apply to exam items in the French language.

For the most part, the LLQP Exam items were deemed to be of acceptable quality. They are all focused, clear, and based on well-developed scenarios. A few minor issues identified in the items include the use of close-ended items that require a “yes” or “no” response, the use of a personal pronoun “you” in the stem, and lack of consistency in the use of professional titles.

There are a few “yes”/”no” items on the LLQP exam. These items have a close-ended question in the stem that requires a “yes” or “no” answer. The answer options, however, go beyond a simple “yes” or “no” and add the reason why a specific action is to be taken. The question does not inquire about the reason; yet the reason is provided in the answers. The “yes/no” items are problematic because there is no perfect alignment between their stem and answer options.

There are also some items on the LLQP exam that use a personal pronoun “you” in the stem. In order to avoid a situation where the candidate personalizes a question or responds from their own unique perspective, an item should refer to an agent in the third person singular or use a professional title.

Finally, there is lack of consistency in the use of professional titles on the exam. For example, the job titles “insurance representative,” “advisor,” and “specialist” are used interchangeably. It would be beneficial to review all exam items for consistent use of terminology.

Exam Fairness

Ensuring fairness for all examinees is an essential goal of all exam programs. Fairness can be defined as giving all examinees an equal opportunity to demonstrate their true standing on the construct that is being measured by the exam. In the context of item development, fairness refers to the absence of any offensive, stereotypical, or unfamiliar content that may distract examinees and prevent them from

demonstrating their "true" competence. Common fairness issues identified in items include the following:

- Offensive language;
- Offensive content;
- Emotionally provocative content;
- Portrayal of gender/racial stereotypes;
- Unequal referrals to men and women;
- Content unfamiliar to a group (e.g., acronyms and abbreviations that may not be familiar to all groups of examinees);
- Vocabulary unfamiliar to a group (e.g., low-frequency or complex words in English or French).

If there is a reason to believe that exam items have content (i.e., words, phrases, or sentences) that may be differentially familiar to different groups of examinees, it will be important to conduct an item sensitivity review for all items on the exam. According to *Standard 3.2*, it is the responsibility of an exam developer to ensure that exam scores are not affected by examinee characteristics that are unrelated to the purpose of the exam.

> ***Standard 3.2*** Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics.

While CISRO does not conduct formal item sensitivity reviews as part of item development, measurement and evaluation specialists and item reviewers check items for clarity and simplicity of the language used. According to the *Standards,* "the level of language proficiency …required by the test should be kept to the minimum required to meet work and credentialing requirements and/or to represent the target construct(s)" (p. 64). In other words, the LLQP exam items should be written in the language that is used by life insurance agents on the job. Apart from reviewing exam questions for complex language, CISRO should also examine them for other types of insensitive content.

Summary and Recommendations

- ***Training and Selection of Item Writers and Reviewers***. In general, basic item development processes are followed. CISRO provides item writers and reviewers with appropriate training and support to assist them in developing multiple-choice items that meet exam specifications and adhere to the general principles of writing multiple-choice items. The instructions received by item writers and reviewers are clearly documented, as is the general process of item development. The requirements of *Standard 4.8* are generally met.

- ***Adherence to Item Writing Principles.*** The psychometric review of English and French items revealed that they are written in accordance with psychometric standards (see Haladyna et al., 2002). The items are of good quality, and do not require any substantial revisions. Yardstick identified a few minor issues that may be considered by CISRO as the organization further develops its item drafting guidelines. The use of items that require a "yes" or "no" answer is not recommended. The personal pronoun "you" in items should be replaced with a pronoun in the

third person singular (i.e., “he” or “she”) or with a relevant professional title. It may be worthwhile to review all exam items for consistent use of terminology.

- ***Exam Fairness*.** It is the responsibility of an exam developer to provide all examinees with an equal opportunity to demonstrate their competence on the exam. The LLQP Exam items are currently reviewed for clarity and simplicity of language. Apart from reviewing exam questions for complex language, CISRO should also examine them for other types of insensitive content.

Stage III –Exam Assembly

For Stage III, Yardstick investigated how exam forms were created and to what extent they conform to exam blueprint parameters. Statistical analyses were conducted to establish the reliability of all exam forms and determine the psychometric properties of items.

Assembling questions into an exam for operational administration is a critical step in exam development. Clear documentation of the process used to assemble the final operational exam forms and the extent to which exam content adheres to exam specifications provides evidence to support the validity of exam score interpretations.

Exam Assembly

When assembling exam forms, CISRO uses competency weights from the exam specifications and theoretical indices of item difficulty to ensure that exam forms have the appropriate breakdown of competencies and similar difficulty levels.

Of central importance in exam validation is demonstrating the representation of content domains. *Standard 4.12* recommends that exam developers document the process used to assemble the final operational exam. Specifically, they are advised to provide a description of the alignment of exam content with exam specifications. This information will support the conclusion that exam performance of examinees reflects their competence in content domains outlined in exam specifications.

> ***Standard 4.12*** Test developers should document the extent to which the content domain of a test represents the domain defined in the test specifications.

CISRO provided Yardstick with a document that contains the mapping of exam items onto competencies from the exam specifications. As can be seen from this document, alternate exam forms for all four exams conform to the exam specifications, which serves as evidence in support of *Standard 4.12*.

The document titled *LLQP Exam Monitoring: Transitional and Ongoing Process* provides criteria for exam assembly. The focus of exam assembly is on meeting exam specifications while simultaneously creating forms that are balanced in terms of statistical and content characteristics. The document differentiates between current exam assembly procedures (i.e., the ones which apply to the exam during the so-called transitional period) and those, which will apply to the exam in the future once enough data on exam items is collected (i.e., maintenance period).

At present, the LLQP exam forms have an approximately equal number of words and an even distribution of items that require calculations, items with longer cases or response options, and items that address specific elements of a competency. Also, in the future, care will be taken to balance exam forms in terms of sex of the person mentioned in the item, this person's profession, and socio-demographic characteristics.

During the transitional stage of exam implementation, the difficulty levels of alternate exam forms are balanced using theoretical indices of item difficulty. A theoretical index of item difficulty is an estimate of the percentage of examinees answering the item correctly. Item writers and reviewers provided such

estimates for all items in the LLQP exam item bank.

During the transitional period, only items with theoretical indices of difficulty of 40% or above and those with discrimination indices above 0.15 are used in exam assembly. The theoretical indices of item difficulty are subjective, which makes them poor indicators of item quality as compared to statistical indices of item difficulty. The latter are based on exam data, and when computed on large samples, serve as reliable indicators of item difficulty.

At present, CISRO uses theoretical indices of item difficulty to create new exam forms and replace poorly-performing items on the existing exam forms. As more exam data is collected for the LLQP exam, CISRO intends to switch to using statistical rather than theoretical indices of item quality for exam assembly. This intention is outlined in the document *LLQP Exam Monitoring: Transitional and Ongoing Process.*

In the future, the criteria for item quality will become slightly more stringent. Only items with difficulty indices between 30% and 85% and those with discrimination indices above 0.20 will be selected for the exams. Also, the guidelines for exam assembly will include the breakdown of items by difficulty level (low: 10% of questions with p-values of 30-49%; average: 60% of questions with p-values of 50-69%; and high: 30% of questions with p-values of 70-85%).

Despite these positive developments, the selection of brand new items for the exams will still be driven by theoretical indices of item difficulty. To remedy this problem, it is recommended that CISRO pre-test new items before they appear on the operational exams. New items can be added as experimental items to the operational exams. Alternatively, they can be pre-tested using a sample of newly licensed life insurance agents.

Equivalence of Exam Forms

Multiple exam forms are often developed to address security issues, such as cheating. When alternate exam forms are used, it is very important to ensure exam score equivalency. Score equivalency implies that examinees would obtain the same scores regardless of the alternate form they complete. Exam scores on alternate exam forms are interchangeable only when these forms were built to the same content and statistical specifications. According to *Standard 5.12*, an exam developer must provide a clear rationale and supporting evidence regarding the extent to which alternate exam forms are comparable in terms of content and statistical information.

> ***Standard 5.12*** A clear rationale and supporting evidence should be provided for any claim that scale scores earned on alternate forms of a test may be used interchangeably.

CISRO meets *Standard 5.12* in the area of content exam specifications. There is documented evidence that three alternate exam forms meet exam specifications in terms of competency representation. At the same time, there does not seem to be any documentation providing statistical information for these alternate exam forms. Given their potential for bias, theoretical indices of item difficulty do not provide solid evidence of equivalence of exam scores on the alternate exam forms. It is recommended that CISRO collect and document information on how alternate exam forms perform in the same exam administration and across multiple exam administrations to be able to claim that exam scores on these

exam forms are equivalent.

Exam Translation and Adaptation

When an exam is administered in multiple languages, it is important that scores on different language forms have the same meaning. According to the *Standards*, simply translating an exam from one language into another does not ensure that the translated exam has the same content and difficulty as the original exam or that it has similar reliability and validity. The *Standards* encourage an exam developer to investigate the validity, reliability, and score comparability of exam forms in different languages. Specifically, *Standards 3.12* and *7.6* state that, when an exam is available in more than one language, the methods used for exam translation and adaptation should be described in detail. Also, evidence of exam score reliability and validity should be provided.

> ***Standard 3.12*** When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for intended use.

> ***Standard 7.6*** When a test is available in more than one language, the test documentation should provide information on the procedures that were employed to translate and adapt the test. Information should also be provided regarding the reliability/precision and validity evidence for the adapted form when feasible.

While CISRO has a robust process for exam translation and adaptation in place, this process is not well documented. Yardstick obtained relevant details through verbal communication with CISRO. Based on that communication, it is clear that exam items are written in any of the two official languages and then translated by AMF translators. Later, a subject matter expert reviews items against the exam specifications and exam preparation manuals in both languages. Another party, called linguistic advisors, collaborates with item writers, translators, and reviewers, on item translation. All the parties receive instructions to avoid using, among others, technical terms, which are not explained in the preparation manuals for the exam, regionalisms and the jargon, units of measurement that are specific to one cultural context. As the last step, measurement and evaluation specialists re-read items in both languages to ensure item readability and clarity in both languages. It can be concluded that CISRO uses the professional judgment of subject matter experts and translators as evidence of exam form equivalence.

The same *Standard 3.12* recommends that an exam developer conduct statistical analyses of different language forms to see if they have similar reliability. Equivalent exam forms are expected to have similar exam- and item-level characteristics. The latter implies similar item p-values and corrected point biserial correlations. It is unknown if CISRO has a process in place for comparing exam-and item-level statistics of English and French exam forms. Such process is recommended. If sample sizes permit, a more comprehensive Differential Item Functioning (DIF) analysis could be conducted using item response theory (IRT) methods.

Exam Form Equating

Although the LLQP exam forms are built to the same content and statistical specifications, they may differ in their actual difficulty levels, creating the need for equating. To maintain the meaning of scores across exam administrations, the LLQP exam is equated using an anchor exam design, which involves a set of common (anchor) items that are placed on all alternate exam forms. According to *Standard 5.15,* the anchor set is supposed to represent a mini-exam that matches the total exam in content and difficulty level. Although there is no standard in terms of an exact number of anchor items needed, it is recommended that the anchor set consist of 20-30 items.

> ***Standard 5.15*** In equating studies that employ an anchor test design, the characteristics of the anchor test and its similarity to the forms being equated should be presented, including both content specifications and empirically determined relationships among test scores. If anchor items are used in the equating study, the representativeness and psychometric characteristics of the anchor items should be presented.

There are 20 items on the *Ethics Common Law Exam*, and 30 items on each of the other three modular exams, including *Life Insurance Exam*, *A&S Insurance Exam*, and *Segregated Funds Exam*. Each alternate form of the LLQP exam has four anchor items, one per each competency component of the exam specifications. There are two observations that could be made about the anchor set used by CISRO. Firstly, the number of items in the anchor set is very small. While all competency components are represented in the anchor set, they do not seem to be weighted in accordance with the exam specifications. It is also not clear if CISRO considers item psychometric properties when creating an anchor set. Since the quality of equating depends on the extent to which an anchor set reflects the total exam in terms of content and psychometric properties, CISRO is recommended to increase the length of the anchor set and use item characteristics for selection of anchor items. Yardstick is aware of the fact that, given the short length of the LLQP exam forms, adding items to the anchor set may not be feasible. An alternative recommendation would be to choose another method of exam equating.

There is no documentation surrounding the process of comparing performance of alternate exam forms after exam data is collected. It is assumed that CISRO evaluates statistical performance of the anchor set across exam forms. Yet in accordance with *Standard 5.17*, direct evidence of score comparability of alternate exam forms must be provided.

> ***Standard 5.17*** When scores on tests that cannot be equated are linked, direct evidence of score comparability should be provided, and the examinee population for which score comparability applies should be specified clearly. The specific rationale and the evidence required will depend in part on the intended uses for which score comparability is claimed.

Summary and Recommendations

- ***Exam Assembly.*** In accordance with the *Standards,* the LLQP exam is assembled using competency weights from the exam specifications. The alignment of exam content with the exam specifications is well documented.

  When assembling alternate exam forms, CISRO balances their difficulty level using theoretical indices of item difficulty provided by subject matter experts. Since these indices are inherently subjective, they may under- or over-estimate the actual difficulty of items. Statistical indices of item difficulty are much more reliable indicators of item quality because they are based on objective data derived from a large sample of examinees.

  It is recommended that CISRO pre-test new questions to obtain statistical indices of item difficulty and discrimination and use both of those indices for exam assembly. CISRO could pre-test new items by putting them on the operational exam forms or by running a formal pilot test with newly licensed life insurance agents.

- ***Equivalence of Exam Forms.*** Alternate exam forms are considered equivalent when they were built to the same content and statistical specifications. CISRO has evidence to support the claim that alternate forms of the LLQP modular exams are similar in content. However, there is no documented statistical information to support the equivalence of exam forms. It is recommended that CISRO collect and record information on how alternate exam forms perform in the same exam administration and across multiple exam administrations. For more information on exam form equivalence, see Section V that describes the results of statistical analyses for alternate exam forms completed by Yardstick.

- ***Exam Translation and Adaptation.*** Since the LLQP exam is administered in both official languages, it is important to establish that scores on English and French exams have the same meaning. CISRO uses the professional judgment of several subject matter experts and translators to ensure that different language forms are equivalent in content. In addition, CISRO should consider empirical investigations of English and French exams to see if they are similar in terms of exam reliability and item statistics. For instance, item statistics could be compared across languages, or a Differential Item Functioning (DIF) analysis could be conducted. Both content review and empirical investigations of English and French exam forms should be documented in detail to provide CISRO with tangible evidence of exam form equivalence.

- ***Exam Form Equating.*** The requirements of the Standards for exam equating are partially met. CISRO uses an anchor set exam design to equate alternate forms of the LLQP modular exams. Ideally, an anchor set should reflect the total exam in terms of content and psychometric properties. While the LLQP exam forms have anchor items for each competency element, the number of anchor items is too small for a meaningful exam form comparison. Since the quality of exam equating depends on the extent to which an anchor set reflects the total exam in terms of content, it is recommended that CISRO put no fewer than 25% of anchor items on any given form and ensure that those items represent a mini-exam that matches the total exam in content, difficulty and discrimination power.

A significant increase in the number of anchor items on the exam will increase an overlap between exam forms. If such overlap is not desirable, another equating method should be considered. For example, alternate exam forms could be equated by using item passmarks obtained in standard setting. Standard setting is discussed in detail in Section VI of this report. Unlike item statistics, item passmarks do not vary with the sample, and thus, provide a good way of calibrating items in terms of their difficulty level. Item passmarks can be used during exam assembly to put together alternate exam forms with the same passing standard.

Stage IV – Exam Administration, Scoring, and Reporting

During Stage IV, Yardstick reviewed exam administration conditions for the LLQP exam, including directions for examinees, exam administration personnel responsibilities, exam security procedures, as well as the testing environment and exam administration process. Scoring and reporting procedures were also evaluated.

The accuracy and usefulness of exam scores depends on the extent to which exam administration, scoring, and reporting procedures are standardized. The standardization of these procedures ensures that all examinees have an equal opportunity to demonstrate their competencies and no one has an unfair advantage. Whenever there are differences in exam administration, scoring, and reporting processes and procedures, the validity of exam score interpretation is compromised.

Exam Administration

Standard 6.1 recommends that an exam developer establish procedures for exam administration and scoring, provide training and support to those responsible for implementing the procedures and provide oversight to ensure adherence to these procedures. For example, instructions to examinees and time limits should be specified and carefully observed. Exam proctors should be trained on how to create standardized administration conditions that support intended uses of score interpretations. They should be aware of their role and responsibilities and know what information they need to communicate to examinees before, during, and after the exam. There should be guidelines on acceptable deviations from standardized exam administration conditions for examinees that require special accommodations. Finally, the exam developer should create a policy regarding re-testing.

> ***Standard 6.1*** Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer and any instructions from the test user.

At present, the administration of the LLQP exam is the responsibility of participating jurisdictions. The exam developer provides jurisdictions with *Practical Exam Administration Guidelines,* a document that sets out general parameters for exam administration, scoring, and reporting. Depending on the policy, *Practical Exam Administration Guidelines* may either provide strong suggestions to jurisdictions (e.g., policy on re-testing) or give them considerable autonomy in decision-making (e.g., policy on special accommodations).

Overall, it appears that the participating jurisdictions have a lot of autonomy over how the LLQP exam is administered. For example, the jurisdictions determine the eligibility of examinees for the LLQP Exam and are authorized to grant exemptions. They decide on the mode of exam administration (on the computer or on paper), examinee identification procedures, and procedures for special accommodations.

The fact that jurisdictions have control over some important components of exam administration creates room for inconsistency in exam administration conditions. It makes it possible for examinees from different jurisdictions to be taking the exam under different conditions. Differences in exam

administration conditions may adversely affect examinee's exam scores, posing a threat to exam score validity and fairness.

For example, in one jurisdiction, examinees may be required to take all exam modules in one day, while the examinees in another jurisdiction may be required to take only one module a day. The examinees in the former jurisdiction are more likely to be tired when taking the last module and as a result, may not do as well on that module as they could. Exam conditions in that jurisdiction may prevent examinees from demonstrating their true knowledge on the exam and adversely affect their scores. In contrast, exam conditions in the other jurisdiction will not affect examinee's scores that way.

The standardization of exam policies and procedures will help CISRO eliminate any construct-irrelevant variance in exam scores. Construct-irrelevant variance refers to the variance in examinee's scores that cannot be explained by the construct measured by the exam. Differences in special accommodations procedures between jurisdictions for example, are potential sources of construct-irrelevant variance in exam scores and thus, should be avoided in testing situations.

Standard 6.4 recommends that an exam be taken in a comfortable environment with minimum distractions, while *Standard 6.5* recommends that examinees be provided with the necessary information about the exam, including practice materials. So, not only should exam conditions be standardized across jurisdictions, examinees should be informed about these conditions prior to the exam.

> ***Standard 6.4*** The testing environment should furnish reasonable comfort with minimal distractions to avoid construct-irrelevant variance.

> ***Standard 6.5*** Test takers should be provided with appropriate instructions, practice, and other support necessary to reduce construct-irrelevant variance.

Standard 8.1 stipulates that an organization should provide examinees with basic information about exam purpose, exam content, and exam administration process to ensure equitable treatment of examinees with respect to access to information. The organization should inform examinees of exam content coverage, including competency area tested and item formats, to help them prepare for the exam. According to *Standard 8.2*, the organization should also provide examinees with information on the testing procedure, exam scoring criteria, the intended use of exam scores, availability of special accommodations, retesting policy, and confidentiality protection. Many organizations choose to put this information in an examinee handbook that is available on the organization's website. In addition to providing valuable logistical information for the examinee, this document is also meant to reduce construct-irrelevant variance related to testing.

> ***Standard 8.1*** Information about test content and purposes that is available to any test taker prior to testing should be available to all test takers. Shared information should be available free of charge and in accessible formats.

> ***Standard 8.2*** Test takers should be provided in advance with as much information about the test, the testing process, the intended test use, test scoring criteria, testing policy, availability of accommodations, and confidentiality protection as is consistent with obtaining valid responses and making appropriate interpretations of test scores.

Standard 6.2 places special emphasis on exam accommodations procedures. Examinees should know which exam accommodations are available to them and what process they follow to obtain them.

> ***Standard 6.2*** When formal procedures have been established for requesting and receiving accommodations, test takers should be informed of these procedures in advance of testing.

The *Guidelines for the Implementation of the LLQP* state that jurisdictions are responsible for adapting the evaluation material for examinees based on the broad principles of fairness specified in the document. It should be noted that only the exam developer has expertise to determine appropriate modifications for the exam, its supporting documentation, or exam administration conditions. These modifications must be done in way that is consistent with the purpose of the exam. In other words, once exam accommodations are provided, the exam still needs to measure what it is supposed to measure in examinees who requested the accommodations.

Standards 6.6 and 6.7 require that organizations make reasonable efforts to ensure the integrity of exam scores and protect the security of exam materials at all times.

> ***Standard 6.6*** Reasonable efforts should be made to ensure the integrity of test scores by eliminating opportunities for test takers to attain scores by fraudulent or deceptive means.

> ***Standard 6.7*** Test users have the responsibility of protecting the security of test materials at all times.

At present, the exam developer leaves it up to the jurisdictions to determine appropriate examinee identification procedures and room set-up. The document titled *Guidelines for the Implementation of the LLQP* recommends that jurisdictions confirm examinee's identity and "provide [exam conditions] that are conducive to examinee's success" (p. 6). It further states that "examinees must be sufficiently comfortable in order to avoid external factors affecting their results" (p. 6). These recommendations are fairly broad since they do not specify how exactly the rooms need to be set up to minimize the opportunities for cheating.

The jurisdictions are also responsible for ensuring the integrity of exam material. The *Guidelines for the Implementation of the LLQP* recommended that exam material and scrap paper be collected at the end of exam session, that the exam be proctored, and that any irregularities during exam administration be brought to the attention of the appropriate authority at the jurisdictional level. However, it is not clear what measures are in place to protect the security of exam material before and after the exam administration, what instructions are given to proctors, and what procedures are followed for examinees suspected of cheating.

Exam Scoring

The quality of exam scoring has direct implications for the reliability and validity of exam score interpretations. Exam scoring must be accurate and consistent across examinees to provide solid support to the decisions made about examinees on the basis of exam scores. The possibility of scoring errors is minimized, if not eliminated, if there are quality assurance processes in place (e.g., answer key verification, double-scoring by an independent party, manual verification of exam scores for a randomly drawn sample of examinees). In fact, *Standards 6.8 and 6.9* require that an organization establish a scoring protocol and have documented quality control processes for exam scoring.

> ***Standard 6.8*** Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.

> ***Standard 6.9*** Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.

The scoring of the LLQP exam is the responsibility of the participating jurisdictions. According to the *Guidelines for the Implementation of the LLQP,* "each jurisdiction is responsible for marking each exam version with the tools of their choice, in compliance with the appropriate answer key provided." (p. 9). It is not clear what quality assurance procedures jurisdictions have in place to confirm the accuracy of exam scoring. Since the LLQP exam could be administered in either computerized or paper-based format, scoring and quality control procedures for computer-based and paper-based exams may vary from jurisdiction to jurisdiction.

Exam Reporting

In reporting exam results to examinees, an emphasis is traditionally placed on ensuring the appropriate interpretation of scores by end users, which in this case are exam examinees and licensing bodies who make decisions about examinees based on their exam scores. Examinees need to know how they performed on the exam and how this performance is evaluated as progress toward the credential. Licensing bodies need to use exam scores only for purposes of making licensing decisions. For example, the licensing bodies should not allow employers to use exam scores for hiring or promotion

purposes. Exam scores should only be used for the purpose for which the exam was developed.

According to the *Standards*, organizations should explain the intended and unintended uses of score report information. According to *Standard 6.10*, score reports should include the description of what an exam covers (e.g., major competency areas), what scores represent (e.g., competence vs. lack of competence), how reliable the scores are, and how exam scores should or should not be used.

> ***Standard 6.10*** When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and how scores are intended to be used.

The responsibility for reporting LLQP exam results lies with the jurisdictions. At present, each jurisdiction determines the type of information regarding exam results to report to examinees, when to report it, and how to report it. It is recommended that the jurisdictions report exam results to candidates in a consistent manner. At minimum, an exam score report should include the number of exam questions by competency area, examinee's exam result (e.g., pass or fail), and the total exam score.

The *Standards* recommend that an organization have a policy on the retention of exam reports and their potential use. According to the *Guidelines for the Implementation of the LLQP*, jurisdictions are recommended to keep "the medium on which examinees transcribe their answers" for five years. No other information is provided on the procedures surrounding the retention of exam information.

> ***Standard 6.14*** Organizations that maintain individually identifiable test score information should develop a clear set of policy guidelines on the duration of retention of an individual's records and on the availability and use over time of such data for research or other purposes. The policy should be documented and available to the test taker. Test users should maintain appropriate data security, which should include administrative, technical, and physical protections

Summary and Recommendations

- ***Exam Administration Procedures.*** Exam scores are more likely to have the same meaning and be more easily interpretable when exam conditions are standardized. The recommendation for CISRO is to regulate all components of the exam administration process by creating a set of comprehensive exam administration policies and procedures that all jurisdictions must adhere to.

- ***Exam Accommodations.*** Yardstick recommends that CISRO specify a range of exam accommodations that are available to examinees and clearly delineate the criteria for requesting and granting these accommodations. Also, the procedures for requesting exam accommodations must be clearly communicated to examinees. Creating a detailed policy on exam

accommodations for all jurisdictions and communicating this policy to examinees will help ensure exam fairness and validity of exam score interpretations.

- ***Exam Integrity*.** Yardstick recommends that the exam developer create a set of detailed exam administration policies and procedures for all jurisdictions to implement. This document will standardize exam administration processes across provinces and contribute to greater fairness of exam and validity of exam score interpretations. This document should describe in detail, the procedures to follow before, during and after the administration of the LLQP exam. This document may contain:

    - Instructions prior to the exam
    - Instructions for exam day (which includes verbal examinee instructions)
    - Procedures to follow during the writing of the exam
    - Procedures to follow at the end of the written exam
    - Instructions following the exam

  Ideally, this document will stipulate requirements for examinee identification. The jurisdictions should not have any discretion in determining what constitutes an acceptable method of examinee identity verification. This document will also provide instructions to proctors for secure handling of exam material before, during, and after the exam and communicate the protocol for reporting irregularities.

- ***Exam Scoring*.** It is recommended that CISRO put the appropriate scoring and quality assurance procedures in writing and communicate them to the jurisdictions. Exam scoring and quality assurance processes need to be synchronized across jurisdictions. The best way to achieve the standardization of scoring is to centralize it. The exam developer may choose to introduce a double-scoring process or manual verification of answers for a select group of examinees.

- ***Exam Reporting*.** To ensure the validity of exam score interpretations, the jurisdictions should report the same performance information to examinees and do it in the same manner. In other words, the standardization of exam score reporting is warranted.

  Exam score reports should include at minimum the information on exam content coverage, exam result (e.g., pass or fail), and total exam score or exam scores by module. Note that some regulatory organizations opt to provide additional performance feedback to failing examinees by including exam component scores in score reports. Such reporting is not appropriate for the LLQP exam due to the small number of items for each exam component. The lower the number of items counted toward a score, the less likely the score is to be reliable. In general, it is recommended to report examinees' total exam score in an exam score report. The total exam score is more likely to be reliable than exam component scores because it is based on a larger number of items.

Stage V – Exam and Item Analysis

In Stage V, Yardstick evaluated the statistical procedures used by CISRO to evaluate the quality of the LLQP exam as a whole and its items. Yardstick also conducted exam and item analyses for all forms of four modular exams to provide an independent verification of exam and question quality.

The procedures for statistical evaluation of the LLQP exam are described in the following documents: *LLQP Exam Monitoring: Transitional and Ongoing Process, Exam Question Renewal Process* and *Report on Transitional Measures: Implementation of the New LLQP Exam.* This suggests compliance with *Standard 4.7* that encourages exam developers to document the procedures for item testing and selection.

> ***Standard 4.7*** The procedures used to develop, review, and tryout items and to select items from the item pool should be documented.

CISRO evaluates the reliability of exam scores, score variability, as well as item difficulty and item ability to differentiate high- from low-scoring examinees.

Exam reliability is assessed with Chronbach's alpha, a statistical index that indicates how well questions "hang together" when measuring a single, uni-dimensional construct. Chronbach's alpha values range from 0.0 to 1.0, with higher values representing higher exam reliability. In general, Chronbach's alpha values above 0.70 are considered acceptable. For high-stakes credentialing exams, Chronbach's alpha values 0.80 or higher are preferred.

An item difficulty level and its discrimination power are reflected in the item p-value and its corrected point-biserial correlation. P-value is a measure of item difficulty that represents a proportion of examinees answering the item correctly. P-value ranges from 0 to 1.0, with smaller values indicating more difficult items. In general, items with p-values below 0.30 are considered very difficult for examinees, while those with p-values above 0.90 or 0.95 are considered very easy.

The corrected point-biserial correlation represents an association of an item score with the total exam score after subtraction of that item score. The corrected point-biserial correlation reflects the extent to which an item discriminates between low- and high-scoring examinees. It ranges from -1.0 to 1.0, with large positive values indicating items that make a meaningful contribution to the total exam score. A large positive corrected point-biserial correlation means that examinees who got the item right also did well on the exam. Items with corrected point-biserial correlations above 0.15 are considered acceptable.

CISRO is currently evaluating the statistical performance of exam items on a weekly basis. According to the *LLQP Exam Monitoring: Transitional and Ongoing Process* document, when an exam pass rate is below 70%, exam scores for some examinees are adjusted by removing one or two items with low p-values. Once the number of written exams reaches 300, exam forms are reviewed and revised to ensure that they contain items with p-values above 40% and discrimination indices above 0.15.

Standard 4.10 states that an exam developer should use an adequate sample size for psychometric evaluation of items, explain the process by which exam items are screened, and document the results of statistical analyses.

> ***Standard 4.10*** When a test developer evaluates the psychometric properties of items, the model used for that purpose (e.g., classical test theory, item response theory, or another model) should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are screened and the data used for screening, such as item difficulty, item discrimination, or differential item functioning (DIF) for major examinee groups, should also be documented….

In general, CISRO's practices for psychometric evaluation of the exam are aligned with *Standard 4.10*. It is commendable that CISRO uses exam data from large samples to review and revise exam forms. The statistical criteria for item evaluation are described fully, in writing.

There is a concern, however, regarding selective adjustment of exam scores for examinees who took the exam with a low-performing cohort. It is inappropriate to review exam scores only when the pass rates are low since it casts a shadow on the exam scores of those who failed the exam when the pass rates were high. What was the reason why their exam scores were not adjusted by removing poorly performing items?

When the LLQP Exam enters its maintenance period, CISRO will apply new procedures to the statistical evaluation of exam items. The new procedures, outlined in the document titled *LLQP Exam Monitoring: Transitional and Ongoing Process,* include monthly statistical analyses of exam forms to identify poorly-performing items and replace them with better ones. The poorly - performing items are the ones with p-values below 30% or above 85% and those with discrimination indices below 0.20.

The exam form updates will be conducted approximately twice a year. When selecting exam items for new exams, CISRO will prioritize items with high discrimination indices that would allow exam forms to achieve the reliability of 0.70. The new procedures for statistical evaluation of the LLQP exam are reasonable.

CISRO will benefit from documenting the results of all exam and item analyses that will be conducted for the LLQP exam in the future. Tracking the statistical performance of multiple exam forms over time will help CISRO gather evidence to support the validity of exam score interpretations. For example, *Standard 2.3* requires that an exam developer estimate exam reliability and include it in score reports. It is clear that CISRO has relevant information available internally. However, it needs to be compiled after each exam administration.

> ***Standard 2.3*** For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant indices of reliability/precision should be reported.

In order to complete an independent statistical evaluation of the LLQP exam, Yardstick performed exam and item analyses for 12 forms of four modular exams administered to candidates in the spring of 2016. The LLQP Exam has been in use since January of 2016. Between January and May of 2016, exam forms were updated twice to improve the quality of items. Exam updates involved item replacements based on the results of statistical analyses.

Every time an exam form was updated, a new version of that form was created. Table 1 below shows exam versions of the LLQP exam forms, the dates when they were in use, and the average number of candidates per exam form. It also shows which exam versions Yardstick selected for exam and item analyses. To ensure the integrity of results, an exam and item analysis was conducted for one exam version. It was not possible to compile exam data across exam versions.

Table 1. Exam Data for Ethics Common Law, Life Insurance, A&S Insurance, and Segregated Funds Exams (English Only)

Exam Version	Dates of Operational Use	Average Number of Examinees per Exam Form	Included by Yardstick in Analyses? Yes/No
1, 2, and 3	January 4 – March 17, 2016 (Ontario) January 7 – March 7, 2016 (all other provinces)	374	No
4.1, 5.1, and 6.1*	March 18 – May 12, 2016 (Ontario) March 8 – May 4, 2016 (all other provinces)	434	Yes (April 26, 2016)
4.2, 5.2, and 6.2	May 13, 2016 - Present (Ontario) May 5, 2016 - Present (all other provinces)	399	Yes (June 6, 2016)
Total		**1207**	

*Note. After exam data from exam versions 4.1, 5.1, and 6.1 was provided to Yardstick for analyses, the exams remained in use until early May. Consequently, more exam data is available for exam versions 4.1, 5.1, and 6.1 than it is shown in the table above.

Exam and item analyses were performed for two versions of alternate LLQP exam forms (i.e., versions 4.1, 5.1, and 6.1 and versions 4.2, 5.2, and 6.2). Tables 2 through 5 below provide a summary of results for exam versions 4.1, 5.1, and 6.1. These versions correspond to three alternate exam forms for each of the following modular exams: *Ethics Common Law, Life Insurance, A&S Insurance,* and *Segregated Funds.*

Each tables provides estimates of exam reliability (i.e., Cronbach's alpha and standard error of measurement), descriptive statistics for exam scores (i.e., an average, minimum, and maximum score, as well as a standard deviation of scores), descriptive statistics for p-values (i.e., an average, minimum, and maximum p-value, as well as a standard deviation of p-values), and descriptive statistics for corrected point biserial correlations (i.e., an average, minimum, and maximum corrected point biserial correlation, as well as a standard deviation of corrected point biserial correlations).

The results for *Ethics Common Law Exam* suggested that alternate exam forms were similar in terms of the average item difficulty (p-value = 0.73, 0.73, and 0.70 for exam versions 4.1, 5.1, and 6.1, respectively) and the average ability of items to differentiate between high and low-scoring examinees (crpb = 0.21, 0.21 and 0.17 for exam versions 4.1, 5.1, and 6.1, respectively). As seen in Table 2, exam versions 4.1 and 5.1 of the *Ethics Common Law Exam* did not have overly difficult items, while there was one difficult item on exam version 6.1. On average, items on all three alternate exam forms had good discrimination power.

The reliability of an exam depends on its length and the range of exam performance demonstrated by examinees. The shorter the exam and the more narrow the range of exam scores, the lower the exam reliability. The *Ethics Common Law Exam* is a fairly short exam consisting of 20 items. It is not expected to have a high level of reliability. Hence, low reliability coefficients of alternate forms of the *Ethics Common Law Exam* were not surprising.

Table 2. Exam and Item Analysis Results for Ethics Common Law Exam (Versions 4.1, 5.1, and 6.1, English Only)

	Version 4.1	Version 5.1	Version 6.1
Number of Examinees	455	410	423
MCQ (k=20)			
Mean	14.52	14.55	14.08
SD	2.86	2.84	2.60
Min	6	5	3
Max	20	20	20
MCQ p-value			
Mean	0.73	0.73	0.70
SD	0.18	0.20	0.21
Min	0.36	0.39	0.13
Max	0.94	0.98	0.94
MCQ crpb			
Mean	0.21	0.21	0.17
SD	0.07	0.09	0.10
Min	0.07	0.04	0.01
Max	0.36	0.36	0.35
Chronbach's alpha	0.62	0.63	0.53
SEM	1.76	1.73	1.78

Table 3 provides a summary of exam and item analysis results for versions 4.1, 5.1, and 6.1 of the *Life Insurance Exam* forms. All exam forms were similar in terms of the average item difficulty and item discrimination power (p-value = 0.66, 0.66, and 0.67 and crpb = 0.18, 0.21, and 0.27 for versions 4.1, 5.1, and 6.1 respectively). There was only one difficult item on exam version 5.1 (p-value= 0.29) and another one on exam version 6.1(p-value= 0.30). The average item discrimination indices for all three forms were acceptable, and the reliability of each form was in the 0.60s or 0.70s range.

Table 3. Exam and Item Analysis Results for Life Insurance Exam (Versions 4.1, 5.1, and 6.1, English Only)

	Version 4.1	Version 5.1	Version 6.1
Number of Examinees	450	429	402
MCQ (k=30)			
Mean	19.89	19.64	20.18
SD	3.87	4.14	4.77
Min	5	4	6
Max	29	30	29
MCQ p-value			
Mean	0.66	0.66	0.67

SD	0.16	0.18	0.16
Min	0.35	0.29	0.30
Max	0.91	0.97	0.98
MCQ crpb			
Mean	0.18	0.21	0.27
SD	0.09	0.10	0.09
Min	0.00	-0.05	0.12
Max	0.31	0.38	0.50
Chronbach's alpha	0.62	0.68	0.77
SEM	2.39	2.34	2.29

Table 4 provides a summary of exam and item analysis results for versions 4.1, 5.1, and 6.1 of the *A&S Insurance Exam* forms. All exam forms were similar in terms of the average item difficulty and item discrimination power (p-value = 0.70, 0.70, and 0.73 and crpb = 0.22, 0.21, and 0.22 for versions 4.1, 5.1, and 6.1 respectively). There were only two difficult items on exam version 5.1 (p-values = 0.16 and 0.22), and there was one difficult item on exam version 6.1 (p-value=0.24). The average item discrimination indices for all three forms were acceptable, and the reliability of each form was approaching 0.70.

Table 4. Exam and Item Analysis Results for A&S Insurance Exam (Versions 4.1, 5.1, and 6.1, English Only)

	Version 4.1	**Version 5.1**	**Version 6.1**
Number of Examinees	466	445	433
MCQ (k=30)			
Mean	21.04	20.87	21.79
SD	4.12	3.76	3.90
Min	6	9	9
Max	30	29	30
MCQ p-value			
Mean	0.70	0.70	0.73
SD	0.17	0.22	0.17
Min	0.35	0.16	0.24
Max	0.97	0.98	0.97
MCQ crpb			
Mean	0.22	0.21	0.22
SD	0.11	0.09	0.11
Min	-0.13	0.05	-0.02
Max	0.45	0.36	0.40
Chronbach's alpha	0.70	0.67	0.69
SEM	2.26	2.16	2.17

Table 5 provides a summary of exam and item analysis results for versions 4.1, 5.1, and 6.1 of the *Segregated Funds Exam* forms. All exam forms were similar in terms of the average item difficulty and item discrimination power (p-value = 0.62, 0.62, and 0.62 and crpb = 0.26, 0.24, and 0.27 for versions 4.1, 5.1, and 6.1 respectively). There was only one difficult item on exam version 4.1 (p-value = 0.28). The average item discrimination indices for all three forms were acceptable, and the reliability of each form exceeded 0.70.

Table 5. Exam and Item Analysis Results for Segregated Funds Exam
(Versions 4.1, 5.1, and 6.1, English Only)

	Version 4.1	Version 5.1	Version 6.1
Number of Examinees	443	442	409
MCQ (k=30)			
Mean	18.47	18.56	18.53
SD	4.93	4.71	5.08
Min	7	4	6
Max	30	30	30
MCQ p-value			
Mean	0.62	0.62	0.62
SD	0.15	0.14	0.15
Min	0.28	0.38	0.34
Max	0.91	0.91	0.83
MCQ crpb			
Mean	0.26	0.24	0.27
SD	0.09	0.09	0.09
Min	0.04	0.04	0.07
Max	0.41	0.41	0.41
Chronbach's alpha	0.76	0.73	0.78
SEM	2.42	2.45	2.38

Tables 6 through 9 below provide a summary of statistical results for exam versions 4.2, 5.2, and 6.2. These exam forms are currently in use.

The results for current versions of the *Ethics Common Law Exam* forms are very similar to those for the previous exam versions. Three alternate exam forms are similar in terms of average item difficulty and item discrimination power (p-value = 0.71, 0.72, and 0.70 and crpb = 0.21, 0.24 and 0.20 for exam versions 4.2, 5.2, and 6.2, respectively). None of the forms has very difficult items, which is an improvement over the previous exam versions. On average, the items have good discrimination power. The reliability of the current exam versions is similar to those of the previous ones. It should be noted that the reliability of form 6.2 has improved.

Table 6. Exam and Item Analysis Results for Ethics Common Law Exam
(Versions 4.2, 5.2, and 6.2, English Only)

	Version 4.2	Version 5.2	Version 6.2
Number of Examinees	422	375	395
MCQ (k=20)			
Mean	14.19	14.45	13.96
SD	2.88	2.98	2.98
Min	6	3	3
Max	20	20	20

MCQ p-value			
Mean	0.71	0.72	0.70
SD	0.18	0.20	0.16
Min	0.38	0.36	0.39
Max	0.95	0.97	0.91
MCQ crpb			
Mean	0.21	0.24	0.20
SD	0.08	0.06	0.09
Min	0.02	0.13	0.05
Max	0.36	0.35	0.33
Chronbach's alpha	0.61	0.67	0.60
SEM	1.80	1.71	1.88

Table 7 provides a summary of statistical results for current versions of the *Life Insurance Exam* forms. All three exam forms are similar in terms of the average item difficulty and item discrimination power (p-value = 0.67, 0.66, and 0.66 and crpb = 0.22, 0.22, and 0.22 for versions 4.2, 5.2, and 6.2, respectively). There are no overly difficult items on these exam versions, which is an improvement over the previous exams. The average item discrimination indices for all three forms are acceptable, and the reliability of each form is above 0.70, which seems to be the result of item replacement.

Table 7. Exam and Item Analysis Results for Life Insurance Exam
(Versions 4.2, 5.2, and 6.2, English Only)

	Version 4.2	**Version 5.2**	**Version 6.2**
Number of Examinees	385	392	393
MCQ (k=30)			
Mean	20.11	19.77	19.56
SD	4.33	4.38	4.43
Min	6	6	6
Max	29	29	29
MCQ p-value			
Mean	0.67	0.66	0.66
SD	0.15	0.16	0.16
Min	0.32	0.34	0.32
Max	0.91	0.96	0.98
MCQ crpb			
Mean	0.22	0.22	0.22
SD	0.09	0.11	0.09
Min	0.03	0.00	0.05
Max	0.38	0.40	0.42
Chronbach's alpha	0.70	0.71	0.70
SEM	2.37	2.36	2.42

Table 8 provides a summary of statistical results for current versions of the *A&S Insurance Exam* forms. All exam forms are similar in terms of the average item difficulty and item discrimination power (p-value = 0.71, 0.68, and 0.69 and crpb = 0.26, 0.25, and 0.25 for versions 4.2, 5.2, and 6.2, respectively). There is only one difficult item on exam version 5.1 (p-value = 0.23), and one difficult item on exam version 6.1 (p-value = 0.26). The average item discrimination indices for all three forms are excellent, and the reliability of each exam form is well above 0.70.

Table 8. Exam and Item Analysis Results for A&S Insurance Exam
(Versions 4.2, 5.2, and 6.2, English Only)

	Version 4.2	Version 5.2	Version 6.2
Number of Examinees	424	392	401
MCQ (k=30)			
Mean	21.33	20.45	20.62
SD	4.47	4.44	4.49
Min	6	7	3
Max	30	29	29
MCQ p-value			
Mean	0.71	0.68	0.69
SD	0.16	0.19	0.17
Min	0.35	0.23	0.26
Max	0.98	0.96	0.98
MCQ crpb			
Mean	0.26	0.25	0.25
SD	0.10	0.09	0.11
Min	0.02	0.12	-0.10
Max	0.45	0.38	0.45
Chronbach's alpha	0.75	0.74	0.74
SEM	2.23	2.27	2.29

Table 9 provides a summary of statistical results for current versions of the *Segregated Funds Exam* forms. All exam forms are similar in terms of the average item difficulty and item discrimination power (p-value = 0.63, 0.65, and 0.63 and crpb = 0.26, 0.25, and 0.26 for versions 4.2, 5.2, and 6.2, respectively). There are no overly difficult items on these exams, which is an improvement over the previous ones. The average item discrimination indices for all three forms are excellent, and the reliability of each form exceeds 0.70.

Table 9. Exam and Item Analysis Results for Segregated Funds Exam
(Versions 4.2, 5.2, and 6.2, English Only)

	Version 4.2	Version 5.2	Version 6.2
Number of Examinees	429	356	418
MCQ (k=30)			
Mean	18.82	19.43	18.66
SD	4.92	4.90	5.03
Min	7	6	4
Max	29	30	29
MCQ p-value			
Mean	0.63	0.65	0.63
SD	0.13	0.12	0.12
Min	0.38	0.41	0.38
Max	0.92	0.90	0.83

MCQ crpb			
Mean	0.26	0.25	0.26
SD	0.10	0.09	0.10
Min	0.00	0.08	0.04
Max	0.49	0.41	0.44
Chronbach's alpha	0.76	0.75	0.76
SEM	2.41	2.45	2.46

The results of exam and item analyses conducted for two different versions of alternate exam forms of the LLQP Exam suggest that exam forms are reliable and consist of quality items that are of reasonable level of difficulty for examinees and have a strong ability to differentiate high- from low-scoring examinees. The sample sizes for both sets of analyses are sufficient to make solid conclusions about the quality of items and exams. The results of statistical analyses for the current exam versions are similar to those for the previous ones.

Summary and Recommendations

- ***Statistical Criteria for Exam Evaluation***. CISRO has a transitional process in place for monitoring exam and item performance and evaluating exam and item quality. It is commendable that CISRO implements exam form updates using statistics obtained from large samples. Yardstick's recommendation is to conduct exam and item analyses on a regular basis and adjust exam scores for all examinees in a cohort if necessary. This strategy will help CISRO preserve the meaning of exam scores across examinees. When the LLQP exam enters its maintenance stage, it will be important to document the results of statistical analyses for all exam forms and all administrations.

- ***Results of Exam and Item Analyses.*** Yardstick conducted statistical analyses to evaluate the psychometric properties of the previous and current versions of the LLQP exam forms. The results are similar across exam versions, with the current exam versions having slightly better psychometric properties than the previous ones. Item replacements helped to eliminate overly difficult items.

  The reliability of current exam forms is within an acceptable range given their short length. The items are of moderate level of difficulty for examinees and have a strong ability to differentiate high- from low-scoring examinees. The alternate forms of all four exams have similar levels of exam reliability and average item statistics, which serves as evidence of their equivalence. It can be concluded that alternate exam forms are comparable not only in terms of content, but also in terms of their psychometric properties.

Stage VI –Standard Setting

During Stage VI, Yardstick reviewed the documentation describing the process used to set a standard on the exam.

Standard setting in the credentialing industry refers to the process of establishing a performance standard (i.e., cut score) on the exam. On credentialing exams, the performance standard divides the score range to partition the distribution of scores into two performance categories – passing and failing examinees. Passing examinees score at or above the standard and are considered competent because they demonstrated knowledge or competencies required for safe and effective practice. Failing examinees score below the standard and are considered as those lacking competence because they failed to demonstrate the required knowledge or competencies.

Standard setting is a critical step in the development and use of credentialing exams because it enables professional associations and regulatory bodies to make appropriate decisions about examinees. Standard setting produces a performance standard that serves as important evidence of validity of exam score interpretation.

There are several testing standards that govern standard-setting processes in credentialing organizations. *The NCCA Standard 17* states that the procedure used to establish performance standards must be based on generally acceptable measurement principles that are consistent with the purpose of the exam. It also recommends that the standard-setting process be documented in sufficient detail to allow for replication, including descriptions of the procedure and outcomes. *Standard 5.21* dictates that, when the standard is used to classify examinees into distinct performance categories, such as passing and failing, the rationale for standard setting procedures must be provided and the procedure should be documented clearly and in sufficient detail.

> ***Standard 5.21*** When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Based on the documentation provided by CISRO, it appears that the passmark for the LLQP Exam was set using a normative procedure. There are two broad categories of standard setting procedures: normative and absolute (or criterion-referenced). Normative standard setting procedures rely on an arbitrary passmark that was set based on the performance of a sample of examinees. These procedures are used in personnel selection where the number of examinees applying for jobs exceeds the number of positions available. Absolute standard setting procedures involve a passmark that is based on the criteria for acceptable professional practice.

The technical report prepared by a third-party consulting firm in 2001 describes a normative procedure to standard setting for the LLQP Exam. However, it does not provide a rationale for selection of this procedure over the absolute one to set a passmark. Absolute standard setting procedures are by far preferred in the credentialing industry because they take into consideration the properties of an exam, the characteristics of the examinee population, and most importantly, the level of performance required to ensure competent practice.

The technical report on standard setting provided three options for calculating the passmark using pilot test data. Yet, it is not clear how the pilot test was conducted, who were the pilot test participants, which exam forms they completed, which group was chosen as a reference group and how this group performed on the exam. Additionally, the recommended passmark was not documented in the final report.

Through email communication, CISRO indicated that the LLQP Exam passmark was set using pilot test data from the group of life insurance professionals who had two years of experience in the industry. The passmark was derived from the lowest exam score of a participant in that group by subtracting a test band score from that score. As of January of 2016, CISRO began to apply the same "historical" passmark to all modular exams. In order to pass the exam, an examinee must successfully complete all modular exams. Also, during the transitional period, CISRO seems to be adjusting the passmark down by one or two points during the weeks when the pass rate falls short of 70%.

Given that the LLQP Exam is a licensing exam that aims to identify competent life insurance professionals, it is recommended that its passmark reflect a standard of proficiency established by experts. This calls for the use of an absolute standard setting procedure where a passmark is based on the level of performance required for safe and competent practice.

There are several concerns associated with the use of a normative procedure for standard setting on credentialing exams. Firstly, normative standards guarantee that some examinee(s) will fail regardless of their level of demonstrated knowledge. If the standard is set at one standard deviation below the mean, it guarantees that 16% of the examinee population will fail the exam regardless of what they know. In the case of the LLQP Exam, it is not clear why performance of the lowest-scoring examinee in the group of life insurance professionals with two years of experience was chosen as a reference point for standard setting. The choice is particularly troublesome given that the pilot test participants may not represent the target population of examinees taking this exam due to the lack of relevant training.

In regard to the standard setting procedure that is appropriate for a licensing exam, Yardstick recommends the modified-Angoff method. There are various procedures available for setting performance standards on exams. The Angoff procedure and its derivatives is one of most commonly used approaches to setting performance standards on multiple-choice credentialing exams. According to the 2008 standard setting survey of organizations with the NCCA accredited certification programs, 75% of organizations used the modified-Angoff procedure. When asked about critical factors that influenced their choice of the standard setting procedure, the organizations mentioned the reliability of the procedure, its acceptability in regard to the NCCA or ISO/IEC accreditation standards, and the ease of use.

Another concern related to the use of the normative standard setting procedure is its lack of sensitivity to changes in examinees' competence levels over time, which can create groups of credentialed professionals with various levels of competence. Additionally, if competence requirements in the industry increased in the past decade, the current passmark on the LLQP Exam would not reflect those changes because it has never been tied to any performance standard. The lack of explicit connection to the entry-level standard of proficiency is the reason why the current passmark cannot support the validity of inferences drawn from exam scores.

Once an exam passmark is established, it is treated as a fixed value until a group of policy makers decide that it is no longer appropriate. The decision to change the passmark is never taken lightly. It is made by the group of policy makers using a standard error of the mean of standard setting ratings. For example, the abovementioned group may decide to raise or lower an exam passmark by adding or

subtracting one or two standard errors of the mean of standard setting ratings. The standard error of the mean indicates the amount of change in the passmark expected upon the replication of the standard setting study. Note that it is impossible to make a valid passmark adjustment without completing the standard setting study first.

Policy makers must have a compelling reason for changing the passmark. For example, they may decide that, in marginal cases, it is preferable to pass than to fail an examinee. Prior to making that decision, the policy makers would compare the consequences of lowering the passmark to those of accepting it as is. For example, lowering the passmark may increase the number of successful examinees who are, in fact, incompetent and obtained a high score on the exam by chance (i.e., false positives). If these examinees pass the exam, they will get certified. The policy makers must weigh in on this risk and decide if it is justified.

According to *Standard 11.16*, an exam passmark should depend on the performance standard (i.e., knowledge and skills required for competent practice) rather than on the exam pass rate. The passmark should not be adjusted to control the number of examinees passing the LLQP Exam.

> ***Standard 11.16*** The level of performance required for passing a credentialing test should depend on the knowledge and skills necessary for credential-worthy performance in the occupation or profession and should not be adjusted to control the number or proportion of persons passing the test.

The NCCA Standard 17 recommends that the certification program evaluate standards of proficiency frequently enough to reflect current practice. Any significant change to the scope of practice or to exam specifications should result in a new standard setting study. The development of the new *Life Insurance Agent Competency Profile* and exam specifications in the recent past makes it clear that a new standard setting study for the LLQP exam is warranted.

Summary and Recommendations

- ***Standard Setting Procedure.*** It is recommended that CISRO conduct a new standard setting study to link the LLQP exam passmark to the absolute performance standard for entry-level life insurance agents in the industry. The use of "historical" passmark by CISRO is problematic for several reasons, and recent change to the scope of practice of life insurance agents necessitates a new standard setting study. Yardstick recommends that CISRO use the modified-Angoff procedure to set a passmark on each module of the LLQP exam. This procedure involves a panel of subject matter experts conceptualizing the expected level of performance for a minimally competent examinee and applying this benchmark to the evaluation of exam questions. For each exam question, standard setters estimate the percentage of minimally competent examinees who will answer the question correctly. Prior to making these judgments, the standard setters receive standardized training and discuss their expectations for minimal competence. Standard setting judgments are made in two rounds, which are separated by discussion with peers and presentation of empirical data indicating item difficulty level. Both discussion and empirical data provide standard setters with feedback that is supposed to increase the accuracy of their judgments. The application of the modified-Angoff procedure will help CISRO establish separate passmarks for each exam form, which will help establish the

equivalence of these forms and support the validity of exam score interpretations. These passmarks will be tied to standards of proficiency in the profession and be reflective of the difficulty of exam questions on a specific exam form.

- ***Documentation of Standard Setting Procedure and Results***. The 2001 technical report on standard setting does not provide sufficient detail on the standard setting process and outcomes. After conducting the new standard setting study, CISRO is advised to document the following information in the standard setting report: 1) rationale for selecting the standard setting method; 2) procedures for selecting and training standard setters; 3) qualifications of standard setters; 4) standard setting procedures; 5) a conceptual description of the target level of performance underlying the standard; 6) data collection activities; 7) the recommended standards; and 8) any adjustments to the standard setting standard made by policy-makers.

- ***Passmark Adjustment.*** Once a passmark is set, it can only be adjusted by a group of policy makers who have a compelling reason to believe that it was set inappropriately or that it is no longer acceptable. A low pass rate can be used to justify passmark adjustment. Yet, once a new passmark is set, it cannot be changed freely to control exam pass rates in the future. Passmark adjustment is defensible only when it is based on statistical data obtained in standard setting.

- ***Passmark Update.*** Once a passmark is established, it needs to be evaluated on a regular basis to see if it still reflects the level of performance required for safe and competent practice. Any change to the scope of practice or body of knowledge for life insurance agents should result in passmark review and adjustment, which may involve a new standard setting study.

Appendix A

Introduction

This addendum provides a commentary and a description of the results of additional data analyses conducted by Yardstick in response to feedback from CISRO stakeholders on the psychometric audit of the LLQP exam. The quality of the LLQP exam is key to obtaining meaningful information about the readiness of candidates to enter the profession of a life insurance agent. Hence, it is important that all relevant parties review and discuss the findings of the psychometric audit.

The following psychometric issues were discussed in this addendum:

1) Impact of the 2016 LLQP modular exams on **exam difficulty and pass rates**;
2) Impact of **exam reliability** on classification decisions about candidates;
3) Need to consider the impact of demographic, socio-demographic, and cultural variables on exam outcomes – **fairness**.

Pass Rates and Difficulty Levels

To investigate the possibility that the introduction of modular exams impacted exam pass rates, CISRO requested that Yardstick compare exam pass rates and difficulty levels for the old and new LLQP exams. The comparison of exam difficulty levels goes beyond the scope of a typical psychometric audit since it involves the old exam, which is no longer in use.

CISRO provided Yardstick with the national pass rates for the 2016 modular exams completed to-date, as well as historical LLQP exam pass rates dating back to 2007. In addition, p-values for items that appeared on three forms of the 2015 LLQP exam in the first 6 months of the year were provided.

Pass Rates

Yardstick compared the national pass rates for the historical LLQP exams to those for the most recent modular exams. Tables 1 through 3 provide the results of exam pass rate comparison.

As shown in Table 1, the average historical pass rates for the old LLQP exam ranged from 64.0% and 74.6%. In 2015, the pass rate of the LLQP exam was 69.2%, which is within range of historical pass rates.

Table A1. Historical Pass Rates for LLQP Exam (2007- 2015)[1]

Year	# Exams Written	Pass Rate
2015	4817	69.2%
2014	11067	72.5%
2013	10180	72.0%
2012	9501	71.9%
2011	8953	74.6%
2010	9250	67.2%
2009	10382	64.0%
2008	9993	64.7%
2007	9982	74.1%

Tables 2 and 3 provide the average pass rates for three exam forms of the 2016 modular exams. For forms 4.2, 5.2, and 6.2, the modular exam pass rates ranged from 66.2% to 84.0% (Table 2), while the current highest and lowest modular exam pass rates are 69.6% and 84.6% (Table 3). Accordingly, there has been a slight upward trend in the national exam pass rates after the introduction of exam forms 4.3, 5.3, and 6.3.

Table A2. Exam Pass Rates for 2016 LLQP Exam (Jan. – June 2016)

Exam Module	Format #	# Exams Written	Pass Rate
Ethics Common Law	4.2	987	82.2%
	5.2	910	84.7%
	6.2	947	85.2%
	Total	**2844**	**84.0%**
Life Insurance	4.2	1137	77.5%
	5.2	1102	71.9%
	6.2	1107	73.1%
	Total	**3346**	**74.2%**
Accident & Sickness	4.2	1195	82.0%
	5.2	1122	79.2%
	6.2	1149	81.1%
	Total	**3466**	**80.8%**

[1] In Table 1, exam pass rates for 2007-2014 are taken from the CISRO website. Exam pass rates for 2015 are calculated for the three exam formats that were in use in all provinces except BC and Quebec for the first 6 months of 2015. These are the three formats that were used for the purposes of comparing the old exam to the new exam in this publication.

Segregated Funds	4.2	1205	64.0%
	5.2	1089	67.7%
	6.2	1183	66.9%
	Total	**3477**	**66.2%**

Table A3. Exam Pass Rates for 2016 LLQP Exam (July 2016)

Exam Module	Format #	# Exams Written	Pass Rate
Ethics Common Law	4.3	456	83.1%
	5.3	425	86.1%
	6.3	432	84.5%
	Total	**1313**	**84.6%**
Life Insurance	4.3	523	78.2%
	5.3	506	78.1%
	6.3	492	73.8%
	Total	**1521**	**76.7%**
Accident & Sickness	4.3	545	80.4%
	5.3	516	80.6%
	6.3	521	83.1%
	Total	**1582**	**81.4%**
Segregated Funds	4.3	560	67.7%
	5.3	525	69.3%
	6.3	509	71.7%
	Total	**1594**	**69.6%**

As can be seen from the summary tables above, with the exception of the *Segregated Funds* Exam, the new modular exam pass rates are slightly higher than the old exam pass rates. In July of 2016, the new *Segregated Funds* Exam was completed successfully by 69.6% of candidates, which is a slight improvement over the 2015 LLQP pass rate of 69.2%.

Also, the average pass rate across all modules of the 2016 LLQP exam is 78.1%. This pass rate is higher than any of the pass rates obtained on the LLQP exam in the past.

In summary, **candidates do better on the 2016 LLQP modular exams than the old LLQP exams.** The comparison of national pass rates between the old and new modular exams showed that conjunctive scoring (i.e., scoring by module) did not affect exam pass rates in a negative way.

Difficulty Levels

Yardstick compared the difficulty levels of the 2015 and 2016 LLQP exams. Specifically, Yardstick compared p-values of items that appeared on the 2015 LLQP exam with those on the two most recent versions of the 2016 LLQP modular exams. The 2015 LLQP exam consisted of 140 items, while the 2016 LLQP modular exams contain 20 or 30 items each. It is worth mentioning that there are significant differences between exam blueprints of the old and new modular exams. The content of any new

modular exam corresponds to only one part of the old exam, which was broader in scope. The old exam contained a number of items that measured competencies from several modular exams simultaneously. Hence, it was not possible to neatly divide the old exam into sections that corresponded to the current modular exams. That was the reason why the old exam was treated as a whole in subsequent analyses. Yardstick recognizes that dividing the old exam into sections by competency would have enabled a more direct comparison of item difficulty levels between the old and new exams.

Table 4 provides a summary of item p-values for three forms of the 2015 LLQP exam. On average, 62% to 64% of candidates answered the 2015 LLQP exam items correctly. The overwhelming majority of items (90% - 95%) were answered correctly by more than one third of candidates. Only 1% or 2% of items on the 140-item exam were considered very difficult (i.e., they were answered correctly by less than one third of candidates). These findings suggest that the 2015 LLQP exam was moderately difficult for candidates.

Table A4. Item p-values for 2015 LLQP Exam

	Form 530	Form 531	Form 532
Minimum p-value	0.23	0.23	0.23
Maximum p-value	0.92	0.97	0.97
Average p-value	0.62	0.64	0.64
SD of p-values	0.16	0.17	0.17
Count of p-values from 0.20 to 0.29	2 (1.4%)	3 (2.0%)	2 (1.4%)
Count of p-values from 0.30 to 0.59	58 (41.4%)	56 (40.0%)	59 (42.1%)
Count of p-values from 0.60 to 0.89	76 (54.3%)	70 (50.0%)	69 (49.3%)
Count of p-values from 0.90 to 1.00	4 (2.9%)	11 (8.0%)	10 (7.1%)
Total count	140 (100%)	140 (100%)	140 (100%)

Note. N (Form 530) = 1689, N (Form 531) = 1623, and N (Form 532) = 1505.

Tables 5 through 8 contain a summary of item p-values for the 2016 LLQP exam forms 4.2, 5.2, and 6.2. These forms were reviewed in the psychometric audit report. On average, 70% to 72% of candidates answered the 2016 *Ethics Common Law* exam items correctly; 66% to 67% answered *Life Insurance Exam* items correctly; 68% to 71% answered *Accident and Sickness Insurance Exam* items correctly; and finally, 63% to 65% answered *Segregated Funds Exam* items correctly. Based on average item difficulty, the new exam was slightly easier than the old one.

The breakdown of the 2016 LLQP modular exam items by p-value was similar to that for the 2015 LLQP exam. Most items were answered correctly by 30% - 90% of candidates. The percentage of items answered correctly by 60-89% of candidates increased compared to the old exam. Also, the distribution of p-values on the new exam was more negatively skewed than on the old exam. **All of these findings suggest that the 2016 LLQP modular exams (forms 4.2, 5.2, and 6.2) were slightly easier for candidates than the old exam.**

Table A5. Item p-values for 2016 Ethics Common Law Exam (Forms 4.2, 5.2, and 6.2)

	Form 4.2	Form 5.2	Form 6.2
Minimum p-value	0.38	0.36	0.39
Maximum p-value	0.95	0.97	0.91
Average p-value	0.71	0.72	0.70
SD of p-values	0.18	0.20	0.16
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	7 (35.0%)	5 (25.0%)	5 (25.0%)
Count of p-values from 0.60 to 0.89	11 (50.0%)	12 (60.0%)	14 (70.0%)
Count of p-values from 0.90 to 1.00	2 (10.0%)	3 (15.0%)	1 (5.0%)
Total	20 (100%)	20 (100%)	20 (100%)

Note. N (Form 4.2) = 422, N (Form 5.2) = 375, and N (Form 6.2) = 395.

Table A6. Item p-values for 2016 Life Insurance Exam (Forms 4.2, 5.2, and 6.2)

	Form 4.2	Form 5.2	Form 6.2
Minimum p-value	0.32	0.34	0.32
Maximum p-value	0.91	0.96	0.98
Average p-value	0.67	0.66	0.66
SD of p-values	0.15	0.16	0.16
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	11 (36.7%)	11 (36.6%)	11 (36.6%)
Count of p-values from 0.60 to 0.89	18 (60.0%)	16 (53.3%)	17 (56.7%)
Count of p-values from 0.90 to 1.00	1 (3.3%)	3 (10%)	2 (6.7%)
Total	30 (100%)	30 (100%)	30 (100%)

Note. N (Form 4.2) = 385, N (Form 5.2) = 392, and N (Form 6.2) = 393.

Table A7. Item p-values for 2016 Accident and Sickness Exam (Forms 4.2, 5.2, and 6.2)

	Form 4.2	Form 5.2	Form 6.2
Minimum p-value	0.35	0.23	0.26
Maximum p-value	0.98	0.96	0.98
Average p-value	0.71	0.68	0.69
SD of p-values	0.16	0.19	0.17
Count of p-values from 0.20 to 0.29	0 (0%)	1 (3.3%)	1 (3.3%)
Count of p-values from 0.30 to 0.59	7 (23.3%)	6 (20.0%)	9 (30%)
Count of p-values from 0.60 to 0.89	21 (70.0%)	19 (63.3%)	17 (56.7%)

Count of p-values from 0.90 to 1.00	2 (6.7%)	4 (13.3%)	3 (10.0%)
Total	30 (100%)	30 (100%)	30 (100%)

Note. N (Form 4.2) = 424, N (Form 5.2) = 392, and N (Form 6.2) = 401.

Table A8. Item p-values for 2016 Segregated Funds Exam (Forms 4.2, 5.2, and 6.2)

	Form 4.2	**Form 5.2**	**Form 6.2**
Minimum p-value	0.38	0.41	0.38
Maximum p-value	0.92	0.90	0.83
Average p-value	0.63	0.65	0.63
SD of p-values	0.13	0.12	0.12
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	12 (40.0%)	10 (33.3%)	10 (33.3%)
Count of p-values from 0.60 to 0.89	17 (56.7%)	19 (63.3%)	20 (66.7%)
Count of p-values from 0.90 to 1.00	1 (3.3%)	1 (3.3%)	0 (0%)
Total	30(100%)	30(100%)	30(100%)

Note. N (Form 4.2) = 429, N (Form 5.2) = 356, and N (Form 6.2) = 418.

In July 2016, CISRO replaced exam forms 4.2, 5.2, and 6.2 with exam forms 4.3, 5.3, and 6.3. Tables 9 through 12 contain a summary of item p-values from the new forms. On average, 71% to 73% of candidates answered the *Ethics Common Law* Exam items correctly; 68% to 69% answered the *Life Insurance* Exam items correctly; 71% to 73% answered the *Accident and Sickness Insurance* Exam items correctly; and 65% to 68% answered the *Segregated Funds* Exam items correctly. The average difficulty level of exam forms 4.3, 5.3, and 6.3 was very similar to that of exam forms 4.2, 5.2, and 6.2.

The breakdown of modular exam items by p-value was similar for two versions of the 2016 LLQP exam. Also, it is noteworthy that CISRO made an effort to eliminate overly difficult items from exam forms. The old LLQP exam had 2 or 3 items that were answered correctly by less than one third of candidates while most of the new LLQP exams have no items like that.

Table A9. Item p-values for 2016 Ethics Common Law Exam (Forms 4.3, 5.3, and 6.3)

	Form 4.3	**Form 5.3**	**Form 6.3**
Minimum p-value	0.42	0.38	0.38
Maximum p-value	0.95	0.97	0.93
Average p-value	0.72	0.73	0.71
SD of p-values	0.18	0.19	0.16
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	8 (40.0%)	5 (25.0%)	5 (25.0%)
Count of p-values from 0.60 to 0.89	10 (50.0%)	11 (55.0%)	14 (70.0%)
Count of p-values from 0.90 to 1.00	2 (10.0%)	4 (20.0%)	1 (5.0%)
Total	20 (100%)	20 (100%)	20 (100%)

Note. N (4.3) = 452, N (5.3) = 421, N (6.3) = 429.

Table A10. Item p-values for 2016 Life Insurance Exam (Forms 4.3, 5.3, and 6.3)

	Form 4.3	Form 5.3	Form 6.3
Minimum p-value	0.49	0.42	0.45
Maximum p-value	0.97	0.96	0.97
Average p-value	0.69	0.69	0.68
SD of p-values	0.15	0.15	0.13
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	12 (40.0%)	10 (33.3%)	10 (33.3%)
Count of p-values from 0.60 to 0.89	16 (53.3%)	17 (56.7%)	18 (60.0%)
Count of p-values from 0.90 to 1.00	2 (6.7%)	3 (10.0%)	2 (6.7%)
Total	30 (100%)	30 (100%)	30 (100%)

Note. N (4.3) = 451, N (5.3) = 439, N (6.3) = 428.

Table A11. Item p-values for 2016 Accident and Sickness Insurance Exam (Forms 4.3, 5.3, and 6.3)

	Form 4.3	Form 5.3	Form 6.3
Minimum p-value	0.32	0.35	0.42
Maximum p-value	0.98	0.97	0.98
Average p-value	0.71	0.72	0.73
SD of p-values	0.17	0.17	0.14
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	5 (16.7%)	7 (23.3%)	4 (13.3%)
Count of p-values from 0.60 to 0.89	23 (76.7%)	20 (66.7%)	22 (73.3%)
Count of p-values from 0.90 to 1.00	2 (6.7%)	3 (10.0%)	4 (13.3%)
Total	30 (100%)	30 (100%)	30 (100%)

Note. N (4.3) = 476, N (5.3) = 441, N (6.3) = 450.

Table A12. Item p-values for 2016 Segregated Funds Exam (Forms 4.3, 5.3, and 6.3)

	Form 4.3	Form 5.3	Form 6.3
Minimum p-value	0.43	0.43	0.50
Maximum p-value	0.90	0.91	0.87
Average p-value	0.66	0.65	0.68
SD of p-values	0.13	0.12	0.11
Count of p-values from 0.20 to 0.29	0 (0%)	0 (0%)	0 (0%)
Count of p-values from 0.30 to 0.59	11 (36.7%)	8 (26.7%)	9 (30.0%)
Count of p-values from 0.60 to 0.89	19 (63.3%)	21 (70.0%)	21 (70.0%)
Count of p-values from 0.90 to 1.00	0 (0%)	1 (3.3%)	0 (0%)
Total	30 (100%)	30 (100%)	30 (100%)

Note. N (4.3) = 487, N (5.3) = 451, N (6.3) = 442.

Based on the comparison of average values and the distribution of item difficulty indices, Yardstick concludes that the 2016 LLQP modular exams are slightly easier than the old exam. This finding is consistent with the one related to the national exam pass rates. It is possible that current exam pass rates, which are slightly higher than before, are explained by the fact that exams are now slightly easier for candidates. That said, exam difficulty is only one of many factors that affect exam pass rates.

Exam Readability

In addition to comparing item difficulty for the 2015 and 2016 LLQP exams, Yardstick examined exam readability. It was assessed using the Flesch – Kincaid grade formula available in the Microsoft Word package. This formula is based on the evaluation of the total number of sentences, words, and syllables in a passage of text. The application of the formula produces a score that corresponds to the U.S. grade level. The more sentences the passage contains, and the longer they are, the higher the grade level. The higher the grade level, the more difficult the passage is to read and comprehend.

The results of the readability analysis of the 2015 LLQP exam are presented in Table 13. The readability of the 2015 LLQP exam forms was approximately Grade 9.

Table A13. Readability Statistics for 2015 LLQP Exam (Forms 530, 531, 532)

Readability Statistics	Form 530	Form 531	Form 532
Flesh - Kincaid Grade Level Index	8.7	8.7	8.8
Passive Sentences %	5	5	5
Sentences per paragraph	2.0	1.9	1.9
Words per sentence	8.0	7.9	8.0
Characters per word	5.0	5.0	5.1

The results of the readability analyses for the 2016 LLQP modular exams (forms 4.2, 5.2, and 6.2) are presented in Table 14 through 17. The readability of the *Ethics Common Law Exam* forms was between Grade 10 and 11. The readability of the *Life Insurance Exam* forms was between Grade 9 and 10. The readability of the *Accident and Sickness Insurance Exam* forms was between Grade 10 and 11. The readability of the *Segregated Funds Exam* forms was between Grade 8 and 9. With the exception of the *Segregated Funds* Exam, the readability indices for the 2016 LLQP modular exams (forms 4.2, 5.2, and 6.2) were one or two grades higher than those for the 2015 LLQP exam.

Table A14. Readability Statistics for 2016 Ethics Common Law Exam (Forms 4.2, 5.2, and 6.2)

Readability Statistics	Form 4.2	Form 5.2	Form 6.2
Flesh - Kincaid Grade Level Index	10.5	10.2	10.8
Passive Sentences %	10%	14%	16%
Sentences per paragraph	1.7	1.8	1.7
Words per sentence	15.8	14.8	15.5
Characters per word	4.9	5.0	5.0

Table A15. Readability Statistics for 2016 Life Insurance Exam (Forms 4.2, 5.2, and 6.2)

Readability Statistics	Form 4.2	Form 5.2	Form 6.2
Flesh - Kincaid Grade Level Index	9.3	9.9	10.1
Passive Sentences %	7%	5%	11%
Sentences per paragraph	2	2	1.8
Words per sentence	14.2	14.8	15.2
Characters per word	4.8	4.9	4.8

Table A16. Readability Statistics for 2016 Accident and Sickness Insurance Exam (Forms 4.2, 5.2, and 6.2)

Readability Statistics	Form 4.2	Form 5.2	Form 6.2
Flesh - Kincaid Grade Level Index	11.0	10.4	10.2
Passive Sentences %	3%	4%	5%
Sentences per paragraph	2.1	2.0	2.2
Words per sentence	13.8	13.2	12.9
Characters per word	5.0	4.9	4.9

Table A17. Readability Statistics for 2016 Segregated Funds Exam (Forms 4.2, 5.2, and 6.2)

Readability Statistics	Form 4.2	Form 5.2	Form 6.2
Flesh - Kincaid Grade Level Index	8.9	9.2	9.3
Passive Sentences %	5%	5%	9%
Sentences per paragraph	1.8	1.7	1.7
Words per sentence	12.4	13.2	12.9
Characters per word	4.8	4.8	4.9

The results of the readability analyses for the 2016 LLQP modular exams (forms 4.3, 5.3, and 6.3) are presented in Table 18 through 21. The readability of the *Ethics Common Law Exam* forms was between Grade 10 and 11. The readability of the *Life Insurance Exam* forms was between Grade 9 and 10. The readability of the *Accident and Sickness Insurance Exam* forms was between Grade 10 and 11. The readability of the *Segregated Funds Exam* forms was between Grade 9 and 10. These readability indices are very similar to those for exam forms 4.2, 5.2, and 6.2.

Table A18. Readability Statistics for 2016 Ethics Common Law Exam (Forms 4.3, 5.3, and 6.3)

Readability Statistics	Form 4.3	Form 5.3	Form 6.3
Flesh - Kincaid Grade Level Index	10.5	10.2	10.8
Passive Sentences %	13%	13%	16%
Sentences per paragraph	1.7	1.8	1.7
Words per sentence	15.8	14.8	15.5
Characters per word	4.9	5.0	5.0

Table A19. Readability Statistics for 2016 Life Insurance Exam (Forms 4.3, 5.3, and 6.3)

Readability Statistics	Form 4.3	Form 5.3	Form 6.3
Flesh - Kincaid Grade Level Index	9.3	10.0	10.1
Passive Sentences %	8%	7%	12%
Sentences per paragraph	2	1.9	1.8
Words per sentence	14	15	15.2
Characters per word	4.8	4.9	4.8

Table A20. Readability Statistics for 2016 Accident and Sickness Insurance Exam (Forms 4.3, 5.3, and 6.3)

Readability Statistics	Form 4.3	Form 5.3	Form 6.3
Flesh - Kincaid Grade Level Index	11.0	10.3	10.1
Passive Sentences %	5%	3%	4%
Sentences per paragraph	2.2	2.0	2.2
Words per sentence	13.9	13.5	13.0
Characters per word	5	4.9	4.9

Table A21. Readability Statistics for 2016 Segregated Funds Insurance Exam (Forms 4.3, 5.3, and 6.3)

Readability Statistics	Form 4.3	Form 5.3	Form 6.3
Flesh - Kincaid Grade Level Index	9.0	9.2	9.3
Passive Sentences %	6%	5%	8%
Sentences per paragraph	1.8	1.7	1.7
Words per sentence	12.4	13.2	12.7
Characters per word	4.8	4.8	4.9

The results of the readability analyses suggest that the old LLQP exam was slightly easier to understand than the new modular exams. Note that this finding does not necessarily mean that the modular exams have high reading requirements, i.e., that a candidate is required to have a reading ability beyond what is expected of a life insurance agent. The new exams are written in the language that Grade 10 or 11 students should be able to understand. What reading ability is expected of life insurance agents? The answer to this question will help to evaluate the readability of the LLQP exam.

According to the *Standards*, the complexity of language used on the exam should correspond to the complexity of language required to fulfill one's job responsibilities.[2] Life insurance agents read documents and communicate with clients on a daily basis. Hence, it is reasonable to suggest that they require the ability to read and comprehend English at least at the high school level.

Ideally, the level of language proficiency required for a profession is derived from a specialized job analysis that focuses on the linguistic requirements of a profession. In Canada, the Centre for Canadian Language Benchmarks is one organization that conducts such job analysis. For example, this organization established language benchmarks for nurses, occupational therapists, and pharmacists. **To our knowledge, there are currently no language benchmarks for insurance professionals. Hence, it**

[2] Standards for Educational and Psychological Testing, 2014, p. 64.

is impossible to make any definitive conclusions about the appropriateness of language used on the LLQP exam.

Exam Reliability

Yardstick was asked to comment on the relationships between the reliability of the 2016 LLQP exam modules and the consistency of classification (pass or fail) of candidates.

The classification decisions about candidates can be correct or incorrect. Correct classification decisions are the ones where candidate's observed and true scores are either above or below the passmark. Incorrect classification decisions include two types of classification errors – false negatives and false positives. A *false negative* refers to a scenario where a candidate's true score falls above the passmark, while their observed score is below the passmark. Consequently, such a candidate fails the exam even though s/he is competent enough to pass it. A *false positive* refers to the opposite scenario wherein a candidate' true score falls short of the passmark but the observed score is above it. Consequently, such a candidate passes the exam even though s/he is not yet competent.

Exam reliability affects the accuracy of classification decisions about candidates (pass or fail). If an exam measures the construct inconsistently (i.e., the amount of measurement error is large), a candidate would be expected to obtain a very different exam score if the candidate were to take the exam again. This would increase the likelihood of a candidate's true score being very different from their observed score (i.e., a classification error). However, in the case of the LLQP exam, classification errors do not present a significant risk for valid score interpretations. There are several reasons for this position.

Firstly, the reliability of three out of four LLQP modular exams (forms 4.2, 5.2, and 5.3) meets conventional standards for reliability (α =0.70). The reliability of the *2016 Ethics Common Law* exam is lower than desired (α =0.60 - 0.70). However, this finding is most likely attributed to the short exam length rather than to item quality.

There are no items on the *2016 Ethics Common Law* exam, which would be overly difficult for candidates. On average, the items are answered correctly by 70% - 72% of candidates. More importantly, item corrected point-biserial correlations are strong. Only a few items (k=2-5) with weak corrected point biserial correlations were found on each exam form. Finally, the exam pass rates for this module in 2016 are rather high (above 80%).

While the current levels of the 2016 LLQP modular exams reliability are acceptable, it is always worth considering how exam reliability can be improved. A more precise measurement would make erroneous classification decisions due to the error of measurement less likely. In case of the LLQP exam, exam reliability could be improved by adding additional items to the exam.

Secondly, classification errors in assessment cannot be completely eliminated. Since no measurement is perfect, there will always be a certain number of misclassified candidates relative to their true scores (both false positives and false negatives). To reduce the likelihood of making incorrect decisions about candidates, it is recommended that CISRO review exam scores of candidates who score near the passmark (i.e., borderline candidates). It is common practice for many organizations to review exam scores of candidates who marginally failed the exam. However, to be fair, the same practice should also apply to candidates who marginally passed the exam. CISRO should have a formal procedure in place for treatment of borderline exam scores. One example of such a procedure is collecting more

information on borderline candidates to better estimate their true score, defined as their expected score over all possible replications of the measurement procedure.

Yardstick did not conduct a formal analysis of decision consistency and accuracy for pass and fail decisions for the 2016 LLQP modular exam because these decisions were not made on the basis of passmarks set using the psychometrically sound and defensible methods (e.g., modified Angoff). Decision consistency and accuracy indices fluctuate as a function of exam score reliability/precision and the location of the passmark. If the passmark does not inspire trust, decision consistency and accuracy indices will be difficult to interpret.

As stated in the psychometric audit report, Yardstick recommends that CISRO conduct a formal standard setting study to obtain an empirical passmark along with the conditional standard error for each exam module. The conditional standard error would then be used to create a confidence interval around the passmark to define the lower and upper limits for passmarks that would be obtained if the same standard setting procedure were repeated many times. The regulatory organization would present both the recommended passmark and the confidence internal to the Exam Advisory Committee, which would set the final passmark for the exam. The Exam Advisory Committee may choose to accept the recommended passmark as it is or adjust it in any direction (up or down) using the standard error information.

Exam Fairness

Exam fairness is a fundamental issue in testing that requires attention of regulators at all stages of exam development and use. It should be noted that exam fairness "has no single technical meaning."[3] At a high level, the Institute for Credentialing Excellence defines fairness as "the degree to which examination candidates are treated similarly and in a reasonable manner."[4] Exam fairness should be considered in item development to ensure that all groups of candidates have equal access to the construct being assessed on the exam. It should also be considered in exam administration and scoring to ensure that all candidates take an exam under the same conditions and receive the same information about their performance.

Exam developers agree that the best way to ensure exam fairness is to build it into development, administration, and scoring processes. Standard 3.0 states that all steps in the exam development, administration, and scoring purpose should be designed in a way as to minimize the impact of factors unrelated to the purpose of an exam on exam scores and promote the validity of exam score interpretations for all groups of candidates.[5]

The notions of exam fairness and exam validity are intertwined. Differential treatment of one group of candidates compared to another group of candidates puts the validity of exam score interpretations for that group in question. Therefore, it is not surprising that many measures aimed at improving the validity of score interpretations will also enhance exam fairness. In the psychometric audit report, Yardstick made a number of recommendations to improve the quality of exam design, administration, and scoring procedures. Many of these recommendations, if implemented, will automatically strengthen exam fairness.

[3] Standards for Educational and Psychological Testing, 2014, p. 49.
[4] Gueorguieva, Koch, & Knapp, 2009, p. 68.
[5] Standards for Educational and Psychological Testing, 2014, p. 63.

For example, in the psychometric audit report, Yardstick discusses the need to standardize exam administration conditions across jurisdictions. If CISRO implements the same exam administration policies and procedures in all jurisdictions, all candidates will be treated in the same manner during exam administration (fairness) and their scores will have the same meaning (validity).

In the report, Yardstick recommends that CISRO conduct sensitivity reviews of exam content to detect words, phrases, or concepts, which may be offensive, stereotypical, or differentially familiar to different groups of candidates. The purpose of this review is to identify aspects of exam content that may prevent certain groups of candidates from demonstrating their competence. Non-native speakers of English or French may be particularly disadvantaged by complex vocabulary or syntax in exam items. While the report does not specifically mention this demographic group, one of the categories of insensitive content featured in the report (e.g., "Vocabulary unfamiliar to a group (e.g., low-frequency or complex words in English or French)" applies to them.

One of Yardstick's recommendations is to collect and document the information on the demographic and socio-demographic characteristics of subject matter experts involved in validity studies, such as the ones underlying competency profile or exam specifications development. This recommendation is aligned with several testing standards cited in the report, as well as Standard 3.3 that states that relevant sub-groups should be included in the validity, reliability, and other preliminary validity studies, such as the pilot test.

The subject matter experts participating in the development and validation of exam questions should be representative of the key demographic characteristics of exam candidates. It is not clear whether this is the case for the LLQP Exam. Relevant information was not available for review at the time of writing the exam audit report. Hence, Yardstick did not explicitly evaluate the demographic composition of validity samples.

To our knowledge, at present, CISRO is not collecting demographic information on exam candidates. Hence, at this point it is impossible to evaluate mean differences in exam scores between different demographic groups, including native and non-native speakers of English or French. In the section on exam translation and adaptation, Yardstick refers to differential item functioning (DIF) analysis as a way to identify items on the English and French exams with statistically significant differences in mean exam scores. Such item analyses could be carried out for native and non-native speakers of English or candidates from other demographic or socio-economic groups to identify if there are any items that perform differently across groups.

Yardstick agreed that it would be useful for CISRO to conduct sub-group analyses of item and exam performance before and after operational use. For example, if CISRO decides to implement Yardstick's recommendation on pilot testing the LLQP exam items, it would be important to collect demographic information on pilot test participants.

It should be noted, however, according to *Standard 3.6*, the presence of sub-group differences in exam or item performance would not indicate lack of fairness.[6] Such differences may exist for legitimate reasons, such as one group having a weaker knowledge of a certain content area than another group. According to Standard 3.6, sub-group differences in item or exam scores should "trigger follow-up studies" to investigate aspects of exam design, content, and format that may contribute to the differential performance of sub-group members.

[6] Standards for Educational and Psychological Testing, 2014, p. 65.

Finally, there is no statistic that can be used to prove that the exam is fair.[7] CISRO should consider and mitigate risks to exam fairness at all stages of exam development, administration, and scoring.

[7] Ziecky, 2002.

References

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.

Berk, R. A. (1996). Standard setting: The new generation (Where few psychometricians have gone before). *Applied Measurement in Education, 9(3),* p. 221.

Chinn, R. N. (2006). Considerations in setting cut scores. *CLEAR Resource Brief.*

Gueorguieva, J., Koch, E. A., & Knapp, D. J. (2009). In Knapp, J., Anderson, L., & Wild, C. (Eds.). *Certification: The ICE Handbook* (pp. 67-92). Washington, DC: Institute for Credentialing Excellence.

Haladyna, T. M., Downing, S. M., & Rodriguez, M.C. (2002). A review of multiple- choice item-writing guidelines for classroom assessment. *Applied Measurement in Education,* 15(3), 309 - 334.

National Commission for Certifying Agencies (2014). *Standards for the Accreditation of Certification Programs.* Washington, DC: Institute for Credentialing Excellence.

Smith, I.L., & Springer, C.C. (2009). In Knapp, J., Anderson, L., & Wild, C. (Eds.). *Certification: The ICE Handbook* (pp. 235-263). Washington, DC: Institute for Credentialing Excellence.

Ziecky. M. (2002). Ensuring the fairness of licensing tests. *CLEAR Exam Review, 12(1),* p. 20-26.