Analyse psychométrique de l'examen du PQAP – Révisée en septembre 2016

À l'intention des Organismes canadiens de réglementation en assurance

Auteure : Natasha Parfyonova, Ph. D.

Psychométricienne principale, Yardstick Inc.

Yardstick
Testing and Training Experts

Table des matières

Sommaire exécutif

Les Organismes canadiens de réglementation en assurance (OCRA) ont confié à Yardstick, un cabinet spécialisé dans l'administration de tests et la formation, le mandat d'effectuer une analyse psychométrique de l'examen du Programme de qualification en assurance de personnes (PQAP). D'après les renseignements recueillis dans le cadre de l'analyse, les processus suivis à l'heure actuelle pour élaborer, faire passer et corriger l'examen du PQAP sont conformes à la majorité des normes régissant l'administration de tests. Les OCRA ont utilisé des méthodes psychométriques adéquates pour définir le contenu et les spécifications de l'examen et pour en élaborer les items. De même, des efforts considérables ont été consentis dans la rédaction, la validation et la traduction des items.

L'évaluation psychométrique des items d'examen en français et en anglais a révélé que ceux-ci ont été rédigés conformément aux principes généraux relatifs à la rédaction de tels items. Il n'a pas été possible de comparer directement le niveau de difficulté de l'ancienne et de la nouvelle version de l'examen du PQAP, car la version la plus récente contient de nouveaux items.

Des analyses statistiques de la performance de l'examen et des items ont été réalisées pour les deux versions anglaises des formats d'examens du PQAP dont l'administration a eu lieu au printemps 2016. Selon les analyses statistiques en question, les formats d'examens ont un niveau de fiabilité auquel on peut s'attendre d'un examen court. En outre, la majorité des items de l'examen ont un pouvoir de discrimination suffisant pour établir une distinction entre les candidats forts et les candidats faibles. Les résultats des analyses statistiques des formats d'examens équivalents sont semblables dans les deux versions de l'examen.

Afin de renforcer le programme d'élaboration et d'administration de l'examen du PQAP, Yardstick a un certain nombre de recommandations à faire; ces recommandations sont présentées en détail dans le présent rapport. Cela dit, il y a deux aspects clés qui, en l'absence d'améliorations, peuvent nuire à la validité de l'interprétation du score de l'examen.

- L'un de ces aspects est l'administration, la notation et la communication des scores de l'examen. Les scores d'un examen sont plus susceptibles d'avoir la même signification d'un candidat à l'autre quand l'examen se déroule toujours dans les mêmes conditions. Autrement dit, il est recommandé de standardiser les politiques et procédures inhérentes à l'administration, à la notation et à la communication des scores de l'examen du PQAP à tous les endroits où des candidats prennent part à l'examen : cette précaution favoriserait le traitement égalitaire des candidats et une interprétation cohérente des scores. En l'absence de politiques et procédures standardisées, il est difficile de garantir l'exactitude de la mesure des compétences des candidats. Yardstick recommande que toutes les juridictions suivent les mêmes politiques et procédures d'administration de l'examen, utilisent le même protocole de notation et communiquent les résultats de l'examen aux candidats d'une manière uniforme.

- L'autre aspect devant faire l'objet d'améliorations est l'établissement d'une norme. L'établissement d'une norme est le processus par lequel est fixé le seuil de réussite à un examen. Le seuil de réussite actuel a été établi il y a plus d'une décennie et, compte tenu de la méthode d'établissement de norme utilisée à l'époque, il n'est associé à aucune norme de performance. Par conséquent, le seuil de réussite ne reflète pas les critères de performance auxquels un représentant en assurance de personnes en début de carrière doit répondre aujourd'hui. Il ne tient pas compte non plus de la difficulté des items de l'examen. Yardstick recommande que les OCRA mènent une nouvelle étude d'établissement de norme pour fixer un

nouveau seuil de réussite à l’examen qui corresponde aux attentes de l’industrie en ce qui concerne les connaissances et les compétences des représentants en assurance de personnes en début de carrière.

La version originale du rapport a été soumise à l’examen des intervenants des OCRA en juillet 2016. En septembre 2016, le rapport a été révisé : de l’information a été ajoutée à la lumière des questions et des commentaires soulevés par divers intervenants.

Introduction

En janvier 2016, Yardstick a reçu le mandat d'effectuer une analyse psychométrique de l'examen du Programme de qualification en assurance de personnes (PQAP) pour les Organismes canadiens de réglementation en assurance (OCRA). L'analyse visait à évaluer l'examen du PQAP, à la lumière des standards de développement des tests, pour déterminer si l'examen remplit adéquatement son objectif : identifier les représentants en assurance de personnes en début de carrière compétents. Au terme de cette analyse, les OCRA pourront déterminer où les processus d'élaboration et d'administration de l'examen se situent par rapport aux normes de développement des tests et quelles améliorations pourraient être effectuées, le cas échéant.

Dans le cadre de l'analyse de l'examen, Yardstick a passé en revue les preuves à l'appui de la fiabilité et de l'équité de l'examen et celles étayant la validité des conclusions à tirer du score obtenu. Les normes d'évaluation de l'examen du PQAP sont principalement issues d'un ouvrage reconnu dans le domaine de l'administration de tests, qui s'intitule *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014). La publication *Standards for the Accreditation of Certification Programs*, de la National Commission for Certifying Agencies (NCCA, 2014), a aussi été utilisée au besoin. Dans le présent rapport, les normes tirées du premier document seront désignées par le terme *normes*, tandis que celles tirées du deuxième seront appelées *normes de la NCCA*.

L'analyse de l'examen a été divisée en six étapes, qui correspondent aux étapes d'élaboration de l'examen. Chaque étape contribue à la qualité de l'examen du PQAP et a une incidence sur l'exactitude des conclusions à tirer du score obtenu. Les six étapes de l'élaboration de l'examen sont brièvement décrites ci-dessous; une description plus élaborée figure dans les parties suivantes du rapport.

ÉTAPE I : Objectif, contenu et spécifications de l'examen

La première étape de l'élaboration de l'examen consiste à définir l'objectif de l'examen, à cibler le contenu approprié et à déterminer la méthode d'évaluation la plus adéquate. Un examen doit porter sur des compétences clés qu'un professionnel doit maîtriser pour exercer ses activités avec rigueur. En outre, ces compétences doivent être évaluées de manière à ce que des conclusions significatives puissent être tirées du score de l'examen. La première étape de l'analyse de l'examen est axée sur les processus relatifs à la définition et à la validation du *Profil de compétences : Représentant en assurance de personnes* et des spécifications de l'examen du PQAP.

ÉTAPE II : Élaboration des items

L'objectif de la deuxième étape de l'élaboration de l'examen est de créer une banque d'items d'examen pertinents. Ces items (ou questions) doivent être élaborés conformément aux pratiques exemplaires et correspondre aux spécifications de l'examen. Dans le cadre de cette analyse, les questions à choix multiple des OCRA ont été évaluées à la lumière des principes inhérents à la rédaction d'items. L'analyse visait aussi à évaluer dans quelle mesure les processus d'élaboration d'items des OCRA permettent à l'organisme de respecter les spécifications de l'examen et de répondre aux besoins généraux liés au programme d'examen.

ÉTAPE III : Assemblage de l'examen

L'assemblage de l'examen désigne les processus sous-jacents à la conception de formats d'examens comparables à partir du contenu de la banque d'items. L'assemblage de l'examen vise à faire en sorte que les formats d'examens concordent avec les spécifications de l'examen et aient un degré de difficulté semblable. Les processus spécifiques d'assemblage de l'examen en place ont été évalués, de même que les preuves confirmant que des formats d'examens équivalents sont conformes aux spécifications de l'examen.

ÉTAPE IV : Administration, notation et communication des scores de l'examen

L'administration de l'examen fait référence aux processus par lesquels les formats d'examens sont soumis aux candidats, tandis que la notation et la communication des scores désignent les processus d'attribution et de transmission des résultats aux candidats. L'analyse vise à déterminer dans quelle mesure les processus d'administration de l'examen, de notation et de communication des résultats sont standardisés, exhaustifs et sûrs.

ÉTAPE V : Analyse d'items et d'examen

L'analyse d'items et d'examen consiste à examiner les propriétés psychométriques d'un examen et des items d'un examen pour déterminer, à l'aide de méthodes statistiques, si l'examen permet de distinguer efficacement les candidats d'après leurs compétences. Un bon examen permet de mesurer les compétences des candidats de façon uniforme. Il permet aussi d'établir une distinction entre les candidats compétents et ceux qui ne le sont pas (ou du moins, pas encore). Cette section du rapport comprend une évaluation des critères statistiques employés par les OCRA pour analyser la qualité des questions de l'examen du PQAP. Elle contient également les résultats des analyses d'items et d'examen réalisées par Yardstick à partir des examens modulaires du PQAP.

ÉTAPE VI : Établissement de la norme

L'établissement de la norme est le processus par lequel le seuil de réussite d'un examen est établi et validé. Pour que le score d'un examen soit significatif, le seuil de réussite doit être établi au terme d'un processus scientifique tenant compte du contenu des items, de leur degré de difficulté et de la performance à laquelle on peut s'attendre d'un candidat minimalement compétent. L'analyse de l'examen évalue le processus d'établissement de la norme employé pour établir le seuil de réussite de l'examen du PQAP et la consignation des résultats.

Les renseignements nécessaires à l'analyse de l'examen du PQAP sont tirés de la documentation écrite fournie par les OCRA et d'entrevues menées avec l'équipe d'élaboration de l'examen des OCRA de l'Autorité des marchés financiers (AMF). Les conclusions et les recommandations présentées dans cette analyse sont basées sur l'information rendue accessible par les OCRA et par l'équipe d'élaboration de l'examen. Les documents suivants ont été utilisés dans le cadre de cette analyse :

- Autorité des marchés financiers (mai 2010).*Le processus de développement des examens de l'Autorité des marchés financiers : une mise en application des meilleures pratiques docimologiques.*

- Autorité des marchés financiers (printemps 2012). *Rapport d'analyse de la profession.*
- CISRO/OCRA (juin 2013). *Profil de compétences : Représentant en assurance de personnes.*
- CISRO/OCRA (juin 2013). *Compte rendu du sondage. Profil de compétences : Représentant en assurance de personnes.*
- CISRO/OCRA (novembre 2013). *Rapport sur les mesures transitoires : Implantation du PQAP*.
- CISRO/OCRA (mars 2014). *Rapport des résultats du sondage : Curriculum du PQAP*.
- CISRO/OCRA (juin 2014). *Programme de qualification du permis d'assurance-vie (PQPAV) : Foire aux questions.*
- CISRO/OCRA (août 2014). *Contrôle de la validité des examens : Implantation d'un PQPAV harmonisé.*
- CISRO/OCRA (septembre 2014). *Life Licence Qualification Program (LLQP): Guidelines for the implementation of the LLQP*.
- CISRO/OCRA (décembre 2014). *Lignes directrices pour la rédaction et la validation : Questions d'examens du PQAP*.
- CISRO/OCRA (2014). *Mesure et évaluation – Feuille de route – automne 2014 : Création et implantation de nouveaux formats.*
- CISRO/OCRA (mai 2015) *Curriculum : Programme de qualification en assurance de personnes (PQAP).*
- CISRO/OCRA (décembre 2015). *Practical exam administration guidelines: Life Licence Qualification Program (LLQP).*
- CISRO/OCRA (n.d.). *Procédures liées au poste d'analyste en mesure et évaluation.*
- CISRO/OCRA (n.d.). *Normes de qualification harmonisées. Experts de contenus : besoins et exigences*.
- Dickson, M. et 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.).
- *Contrôle des examens du PQAP : Processus de transition et de maintenance* (mai 2016).

La version originale du rapport d'analyse psychométrique a été soumise à l'examen des intervenants des OCRA. À la lumière des questions et des commentaires soulevés par les intervenants, le rapport a été modifié : l'annexe A a en effet été ajoutée pour fournir de l'information supplémentaire sur l'examen du PQAP, notamment en ce qui concerne les résultats de la comparaison des taux de réussite et des niveaux de difficulté associés à l'ancienne et à la nouvelle version de l'examen.

Étape I – Objectif, contenu et spécifications de l'examen

Dans le cadre de l'étape 1, Yardstick s'est penchée sur l'objectif de l'examen du PQAP ainsi que sur les processus relatifs à la définition du *Profil de compétences : Représentant en assurance de personnes* et des spécifications de l'examen. Cette analyse avait pour but de déterminer dans quelle mesure le contenu de l'examen concorde avec l'objectif et les spécifications de l'examen.

Pour élaborer un examen et en évaluer les résultats, il faut d'abord et avant tout en énoncer clairement le ou les objectifs; c'est une étape incontournable. Un bon examen a un objectif bien défini qui concorde avec l'interprétation envisagée des scores.

Objectif de l'examen et utilisations envisagées des scores

La *norme 1.1* précise que le concepteur d'un examen doit déterminer clairement le construit que l'examen permet de mesurer, la population à laquelle s'applique l'examen et ce à quoi serviront les scores obtenus. La *norme 1.2* indique quant à elle que le concepteur doit aussi expliquer pourquoi le score d'un examen peut être utilisé aux fins visées.

> ***Norme 1.1*** Les concepteurs de test devraient expliquer clairement la façon d'interpréter et d'utiliser les scores d'un test. La ou les populations pour lesquelles le test a été conçu devraient être clairement délimitées et le ou les construits que le test est censé mesurer devraient être décrits avec précision.

> ***Norme 1.2*** Chaque interprétation et chaque utilisation des scores d'un test qui sont envisagés devraient être justifiés et accompagnés d'un résumé des éléments de preuve et de la théorie qui se rapportent à l'interprétation désirée.

L'examen du PQAP est une norme de qualification à laquelle doivent se conformer tous les Canadiens qui souhaitent obtenir un permis de représentant en assurance de personnes. Seules les personnes ayant suivi un programme de formation approprié peuvent passer l'examen du PQAP. Une fois l'examen réussi, le candidat obtient un certificat et peut soumettre sa candidature pour obtenir un permis en assurance de personnes auprès de son organisme de réglementation provincial.

Comme l'indique le document intitulé *Programme de qualification du permis d'assurance-vie (PQPAV) – Foire aux questions*, l'examen du PQAP a été conçu par les OCRA dans le cadre d'un processus de collaboration avec les organismes de réglementation canadiens dans le domaine de l'assurance. Son objectif est de « protéger les consommateurs en aidant les agents à posséder les connaissances financières requises au sujet des produits d'assurance-vie » (p. 1). De même, selon le document intitulé *Curriculum – Programme de qualification en assurance de personnes (PQAP)*, l'examen du PQAP a pour but d'évaluer la capacité du candidat à adopter une « pratique éthique respectant les droits des consommateurs » (p. 1).

Le construit que l'examen du PQAP vise à évaluer est décrit dans le document *Profil de compétences : Représentant en assurance de personnes*, tandis que les spécifications de l'examen du PQAP reconnues dans l'industrie sont désignées par le terme *curriculum*. L'examen du PQAP comporte quatre examens

modulaires qui ont pour but d'évaluer les compétences du candidat dans les domaines suivants : 1) éthique; 2) assurance-vie; 3) assurance contre la maladie ou les accidents; et 4) fonds distincts. Il vise en outre à évaluer les compétences que doivent posséder les représentants en assurance de personnes en début de carrière.

Comme l'indique le manuel d'élaboration d'examens intitulé *Le processus de développement des examens de l'Autorité des marchés financiers : une mise en application des meilleures pratiques docimologiques*, l'examen du PQAP est un examen à livre ouvert. Autrement dit, les candidats peuvent trouver les informations concernant les réponses aux questions dans la documentation de référence fournie. Cette information contribue à mieux définir l'objectif de l'examen et le construit que l'examen permet d'évaluer.

Profil de compétences

L'une des étapes fondamentales de l'élaboration d'un examen consiste à définir son contenu. Pour assurer la validité des conclusions tirées à partir du score d'un examen, le contenu de l'examen doit refléter le construit qu'il est censé évaluer. Plus le contenu de l'examen est en phase avec le construit, plus il est probable que le score de l'examen puisse être interprété de manière significative.

L'examen du PQAP, par exemple, permet d'évaluer des compétences critiques chez des représentants en assurance de personnes en début de carrière pour veiller à ce qu'ils exercent leur profession de manière éthique et compétente. La validité de l'interprétation du score d'un examen dépend des preuves selon lesquelles les questions de l'examen sont arrimées aux compétences attendues des représentants en assurance de personnes en début de carrière. Ces preuves reposent sur le jugement d'experts de contenu et sont recueillies dans le cadre du processus d'élaboration et de validation de l'examen. Selon la *norme 1.11*, le concepteur d'un examen est tenu de décrire et de documenter les processus qui sous-tendent la définition des compétences à maîtriser par les futurs candidats à l'examen. Il convient également d'énoncer les critères relatifs à la sélection des compétences, tels que l'importance, la fréquence ou l'aspect critique.

> ***Norme 1.11*** Quand l'interprétation du score d'un test s'appuie, en partie, sur la pertinence du contenu du test, les processus suivis pour établir et générer le contenu du test devraient être décrits et justifiés à la lumière de la population de candidats potentiels et du construit que le test prétend mesurer ou du domaine qu'il est censé représenter. Si la définition du contenu échantillonné comporte des critères tels l'importance, la fréquence ou l'aspect critique, ces critères devraient aussi être clairement expliqués et justifiés.

Les renseignements sur les compétences à maîtriser sont tirés des documents suivants : 1) *Profil de compétences : Représentant en assurance de personnes*; 2) *Compte rendu du sondage*; et 3) le manuel d'élaboration d'examens. Le document *Profil de compétences : Représentant en assurance de personnes* a été conçu en 2012 dans le cadre d'un processus en deux volets ayant nécessité des consultations approfondies avec des experts de contenu de cinq provinces : l'Alberta, la Colombie-Britannique, l'Ontario, le Nouveau-Brunswick et le Québec. Trois ateliers d'analyse de la profession se sont déroulés avec des experts des professions suivantes : représentant en assurance de personnes, représentant en assurance contre la maladie ou les accidents et représentant en assurance collective de

personnes. Le processus de validation des compétences consistait en l'administration d'un sondage sur la validation des compétences.

La *norme 1.8* met l'accent sur la nécessité de décrire avec une grande précision les échantillons utilisés aux fins de la validation, y compris l'expérience, les qualifications et les caractéristiques sociodémographiques des participants ayant trait à l'objectif de l'examen. La *norme 1.9* réitère l'importance de préciser quelles sont les procédures de sélection et de formation des experts de contenu. Elle traite notamment des directives qui sont données à ces experts en ce qui a trait à la collecte de données et aux processus inhérents à la prise de décisions. La *norme 7.5* recommande de mettre le tout par écrit.

> ***Norme 1.8*** La composition de tout échantillon de candidats pour lequel on obtient des preuves de validité devrait être décrite avec le plus de détails possible, y compris les principales caractéristiques sociodémographiques et de développement pertinentes.

> ***Norme 1.9*** Quand une validation s'appuie, en partie, sur les opinions ou les décisions d'experts, d'observateurs ou d'évaluateurs, les processus de sélection de tels experts et l'obtention de leurs jugements ou de leurs évaluations devraient être clairement précisés. On devrait indiquer les qualifications et l'expérience des experts. La description des procédures devrait inclure toute formation ou directive donnée aux évaluateurs; elle devrait aussi indiquer si ces experts ont porté leur jugement de façon indépendante et préciser le degré de consensus obtenu par ces derniers. Si les experts interagissent ou échangent de l'information, les mécanismes par lesquels ils ont pu s'influencer les uns les autres devraient être présentés.

> **Norme 7.5** Les documents relatifs au test devraient faire mention des caractéristiques pertinentes des personnes ou des groupes qui ont pris part aux activités de collecte de données associées à l'élaboration ou à la validation de tests (p. ex., données démographiques, statut professionnel, scolarité); de la nature des données fournies (p. ex., données prédictives, données critériées); de la nature des jugements posés par les experts de contenu (p. ex., liens avec la validation du contenu); des directives qui ont été transmises aux participants pour l'accomplissement de leurs tâches dans le cadre des activités de collecte de données; et des conditions dans lesquelles les données relatives au test ont été recueillies au cours de l'étude de validité.

Selon le manuel d'élaboration d'examens, les ateliers d'analyse de la profession visaient à : 1) mettre en lumière les tâches effectuées par les représentants en assurance de personnes; 2) déduire quelles sont les connaissances, les compétences et les habiletés nécessaires à la réalisation de ces tâches; et 3) déterminer les exigences de qualification des représentants en assurance de personnes. Les ateliers se sont déroulés selon la méthodologie standardisée d'analyse des professions approuvée par les gouvernements fédéral et provinciaux. Une description détaillée de cette méthodologie est présentée dans le document intitulé *Rapport d'analyse de la profession.* Ce document contient des renseignements sur les critères de sélection des participants à l'atelier, leurs qualifications, les tâches qu'ils ont accomplies au cours de l'atelier et l'issue des discussions. Il convient de mentionner que les

caractéristiques sociodémographiques et l'expérience des participants à l'atelier n'ont pas été consignées.

À la suite des ateliers d'analyse de la profession, une première ébauche du document *Profil de compétences : Représentant en assurance de personnes* a été créée à l'échelle nationale. Le document en question comprend une liste des « tâches et [...] opérations [...] que peut effectuer, au seuil d'entrée sur le marché du travail, un représentant en assurance de personnes » (c'est-à-dire dans les trois années suivant son entrée en poste) (p. 3).
Le document *Profil de compétences : Représentant en assurance de personnes* est divisé en trois grandes sections : *les blocs*, *les compétences* et *les éléments de la compétence*.

Au terme des ateliers, un sondage a été effectué en ligne pour valider le contenu du *Profil de compétences : Représentant en assurance de personnes* auprès d'un échantillon plus vaste d'experts de contenu. Le sondage, auquel on pouvait répondre dans les deux langues officielles, s'adressait à des personnes occupant différents rôles dans l'industrie : agents, courtiers, représentants d'organismes de réglementation provinciaux, prestataires de cours, dirigeants d'entreprises et professionnels des ressources humaines, entre autres.

En tout, 751 personnes ont répondu au sondage. Comme prévu, les agents (58 %) et les courtiers (29 %) ont été les plus nombreux à y répondre. Il est difficile d'évaluer si les résultats du sondage peuvent être généralisés à l'ensemble des intervenants au pays, les répondants étant répartis de façon très inégale entre les provinces. Le sondage révèle en effet que 72 % des répondants proviennent de l'Alberta, qui n'est pourtant pas la province où travaillent le plus grand nombre de représentants en assurance de personnes. Après avoir mis en parallèle les résultats de l'Alberta avec ceux des autres provinces, les responsables de la définition des compétences ont conclu qu'ils étaient comparables.

Les répondants au sondage ont été invités à porter deux jugements globaux sur le profil de compétences en général (c.-à-d., « Y a-t-il des compétences ou des éléments de la compétence que vous jugez pertinents à la pratique, mais que vous n'avez pas retrouvés dans le profil de compétences? » et « Y a-t-il des compétences ou des éléments de la compétence du profil qui NE sont PAS pertinents à la pratique? »). Plus de 90 % des répondants ont indiqué que le document était complet et 98 % ont affirmé que toutes les compétences étaient pertinentes. À partir des commentaires qualitatifs issus du sondage, des modifications ont été apportées au *Profil de compétences : Représentant en assurance de personnes*.

Spécifications de l'examen

Une fois que le profil de compétences lié à l'examen est établi, il faut convertir ces compétences en spécifications de l'examen. Les spécifications de l'examen constituent une feuille de route détaillée pour l'élaboration d'un examen. Elles servent aussi de preuve de validité du contenu à l'appui de l'interprétation des scores de l'examen.

La *norme 4.1* recommande que les spécifications de l'examen comprennent une description de l'objectif de l'examen, de la population de candidats potentiels, du construit à mesurer et de l'utilisation envisagée du score obtenu. En outre, selon la *norme 15 de la NCCA*, il convient d'indiquer le degré de pratique des candidats (p. ex., débutant, avancé).

Norme 4.1 Les spécifications doivent faire état de l'objectif du test, de la définition du construit ou du domaine mesuré, de la population de candidats potentiels et des interprétations relatives aux utilisations envisagées. Les spécifications doivent comprendre une justification à l'appui des interprétations et des utilisations des résultats du test selon les fins projetées.

Pour harmoniser le processus de qualification en assurance de personnes au Canada, les OCRA ont créé un nouvel ensemble de spécifications de l'examen : ces spécifications portent également le nom de curriculum de l'examen du PQAP. Les spécifications de l'examen regroupent un certain nombre de tableaux d'évaluation qui contiennent des compétences et des éléments de compétence qui doivent être mesurés par l'examen du PQAP. Chaque compétence est évaluée par un examen modulaire, et tous les examens modulaires sont pondérés de façon égale dans le processus d'évaluation. Pour réussir l'examen du PQAP, le candidat doit obtenir au moins la note de passage à chacun des quatre examens modulaires.

Chaque examen modulaire vise à évaluer deux ou quatre éléments de la compétence, lesquels sont pondérés en fonction de leur importance relative pour la protection des consommateurs et de la complexité des concepts qui les sous-tendent. Chaque élément de la compétence est divisé en sous-éléments dont la mesure se limite à une liste de sujets désignée par le terme « contenu ». En outre, les spécifications de l'examen font mention du temps d'administration et de la longueur de chaque examen modulaire.

Le processus inhérent à la définition des spécifications de l'examen est décrit dans le manuel d'élaboration d'examens. Les spécifications de l'examen ont été définies par un groupe composé de spécialistes de la mesure et l'évaluation, de formateurs et de praticiens à l'aide du contenu du *Profil de compétences : Représentant en assurance de personnes*. Les renseignements sociodémographiques et les qualifications des membres du groupe n'ont pas été consignés dans le manuel d'élaboration d'examens. Cela dit, Yardstick sait que les OCRA recueillent des renseignements sociodémographiques sur tous les volontaires qui participent aux activités d'élaboration et de validation d'examen.

Les membres du groupe ont sélectionné les éléments de compétence à évaluer dans le cadre de l'examen et ont déterminé l'importance relative de chaque élément dans la définition d'une compétence (p. 5). La sélection des éléments d'une compétence reposait sur les critères suivants : 1) pertinence par rapport à la mission de protection des consommateurs des OCRA; 2) pertinence des éléments pour des agents en début de carrière; et 3) possibilité de mesurer ces éléments par des questions à choix multiple. Le processus par lequel l'importance relative des éléments d'une compétence a été déterminée n'a pas été documenté. Il semble que les pondérations de la compétence aient été établies d'après le jugement professionnel des experts de contenu.

La *norme 4.2* précise qu'il convient de fournir l'information ayant trait au contenu du test, au nombre d'items proposé, au format des items et au délai imparti pour faire passer le test, de même que les directives données aux candidats et les procédures de notation et de communication des scores.

Norme 4.2 En plus de décrire les utilisations envisagées d'un examen, les spécifications d'un test devraient définir son contenu, préciser le nombre d'items proposé, le format des items, les propriétés psychométriques des items et du test lui-même ainsi que l'ordre des items et des sections. Les spécifications devraient aussi préciser le temps requis pour passer le test, les directives à transmettre aux candidats, les procédures d'administration du test – y compris les variantes permises –, la documentation à utiliser et les procédures de notation et de communication des scores. S'il s'agit d'un examen informatisé, les spécifications doivent comprendre une description des exigences matérielles et logicielles.

Les spécifications de l'examen du PQAP contiennent des renseignements importants sur l'examen : les pondérations attribuées aux différents éléments de compétence, le temps alloué à l'administration de chaque examen et la longueur des examens, par exemple. En revanche, d'autres renseignements clés sont absents. Par exemple, les spécifications ne font pas mention du type d'item (p. ex., questions à choix multiple avec quatre choix de réponses), de la possibilité de passer l'examen dans les deux langues officielles ainsi que des procédures de notation et de communication des scores.

Selon la *norme 15 de la NCCA*, le concepteur d'un examen doit faire mention de l'information suivante dans les spécifications de l'examen :

- Objectif de l'examen;
- Description de la population de candidats;
- Description du construit et du type d'item;
- Aperçu du contenu pondéré;
- Critères d'assemblage de l'examen;
- Exigences relatives à l'administration de l'examen (p. ex., examen informatisé, accès à des documents de référence, utilisation de calculatrices);
- Description générale du plan de notation et de mise en équivalence de l'examen et d'exécution d'analyses psychométriques.

L'étape suivante de l'élaboration des spécifications de l'examen du PQAP consistait à mener un sondage auprès des intervenants. Comme le prescrit la *norme 4.6*, les OCRA et différents organismes de réglementation provinciaux ont invité les intervenants de l'industrie à répondre à un sondage en ligne pour évaluer le caractère approprié des spécifications de l'examen. Le processus d'évaluation et les résultats obtenus sont décrits dans le document *Rapport des résultats du sondage : Curriculum du PQAP*.

La *norme 4.6* préconise de documenter l'objectif de la révision des spécifications de l'examen, le processus selon lequel cette révision a été effectuée, les résultats de la révision ainsi que les qualifications, l'expérience et les caractéristiques démographiques des réviseurs.

> ***Norme 4.6*** Lorsqu'il est nécessaire de documenter la validité des interprétations de scores au test par rapport aux utilisations envisagées, des experts pertinents et indépendants du programme d'administration du test devraient réviser les spécifications de l'examen. Ce processus permet d'évaluer si les spécifications en question concordent avec les utilisations envisagées des scores du test et si elles sont équitables pour les candidats potentiels. L'objectif et les résultats de cette révision ainsi que le processus utilisé devraient être documentés. Les qualifications, les expériences pertinentes et les caractéristiques démographiques de ces experts devraient également l'être.

Les exigences de la *norme 4.6* ont été respectées. Le rapport des résultats du sondage sur les spécifications de l'examen décrit le processus par lequel le lien vers le sondage a été communiqué aux répondants. Il contient aussi de l'information sur les caractéristiques de l'échantillon sondé. Le sondage a été réalisé auprès de 386 répondants, notamment des agents (59 %), des courtiers (23 %) et des personnes exerçant des professions connexes, comme des représentants d'organismes de réglementation provinciaux ou des directeurs des ventes (18 %). La province de résidence était la seule donnée démographique disponible au sujet des répondants. Comme c'était le cas dans le sondage sur la validation des compétences, les répondants de l'Alberta étaient surreprésentés dans l'échantillon (69 %). À l'inverse, ceux de l'Ontario étaient sous-représentés (13 %). Les concepteurs du sondage ont affirmé avoir comparé les résultats globaux de l'enquête avec ceux fournis par les répondants provenant d'autres provinces que l'Alberta, sans avoir décelé de différences. Le rapport des résultats du sondage fait état des questions du sondage et en décrit les résultats en détail.

Résumé et recommandations

- ***Objectif de l'examen et utilisations envisagées du score obtenu.*** L'objectif de l'examen du PQAP, la population de candidats potentiels, le construit évalué et l'utilisation envisagée du score obtenu sont, à l'heure actuelle, énoncés dans différents documents techniques. Pour assurer la conformité aux *normes 1.1 et 1.2*, il est important que les OCRA mentionnent clairement ces informations dans le manuel d'élaboration d'examens et dans les documents accessibles au public (p. ex., un manuel du candidat).

- ***Profil de compétences.*** Des procédures psychométriques adéquates ont été employées pour élaborer le domaine des contenus de l'examen, c'est-à-dire le *Profil de compétences : Représentant en assurance de personnes.* Les procédures en question comportaient des ateliers d'analyse de la profession et un sondage de validation en ligne menés auprès d'un large échantillon d'experts dans les domaines appropriés. Les exigences des *normes 1.8, 1.9, 1.11* et *7.5* sur la documentation du processus inhérent à l'élaboration du profil de compétences ont été respectées. Le processus de déroulement des ateliers d'analyse de la profession et du sondage de validation des compétences est largement documenté par les OCRA. Les *normes* préconisent de consigner les caractéristiques sociodémographiques des échantillons utilisés dans les études de validité. Il convient toutefois de noter que les *normes* ont été publiées aux États-Unis et que certaines d'entre elles ne s'appliquent pas nécessairement au contexte juridique canadien. Il est suffisant de consigner la représentativité géographique des experts, leurs qualifications, leur expérience professionnelle et leurs domaines de spécialité.

- ***Spécifications de l'examen.*** Les OCRA ont employé une démarche psychométrique rigoureuse pour élaborer des spécifications détaillées pour l'examen et assurer leur validation. Les compétences à évaluer dans le cadre de l'examen sont issues du *Profil de compétences : Représentant en assurance de personnes*, et toutes les décisions importantes ayant trait au contenu et à la structure de l'examen du PQAP ont été prises en consultation avec l'industrie. La population de candidats potentiels et le construit à mesurer par l'examen sont décrits de façon très détaillée dans les spécifications de l'examen. Pour être pleinement conformes à la *norme 4.1* et à la *norme 15 de la NCCA*, les OCRA devraient ajouter aux spécifications de l'examen des renseignements sur le type d'item, les critères d'assemblage de l'examen et les exigences relatives à l'administration de l'examen (p. ex., examen informatisé, accès à des documents de référence, utilisation de calculatrices). Il serait aussi approprié d'ajouter une description générale du plan de notation et de mise en équivalence de l'examen et d'exécution de l'analyse psychométrique.

Étape II – Élaboration des items

Dans le cadre de l'étape 2, Yardstick a évalué les processus et procédures employés pour élaborer et réviser de nouveaux items pour l'examen du PQAP. Yardstick a plus particulièrement examiné dans quelle mesure les processus et procédures d'élaboration d'items répondent aux normes régissant l'élaboration et la révision de ceux-ci.

La qualité d'un examen repose sur celle de ses items. Une fois que les spécifications de l'examen ont été définies, il faut élaborer des items conformes à ces spécifications.

Sélection et formation des rédacteurs et réviseurs d'items

On ne saurait trop insister sur l'importance de sélectionner et de former rigoureusement les rédacteurs et les réviseurs d'items. Selon la *norme 4.7*, il est important de documenter la façon dont les rédacteurs et les réviseurs d'items sont sélectionnés, la formation qu'ils reçoivent et la démarche qu'ils suivent pour créer et réviser des questions. Cette information, en plus de donner de la crédibilité au processus d'élaboration d'items, renforce la validité des scores d'un examen.

> ***Norme 4.7*** Les méthodes utilisées pour développer, réviser et expérimenter les items et les choisir à partir d'un regroupement d'items devraient être documentées.

Les exigences auxquelles doivent se conformer les rédacteurs et les réviseurs d'items sont décrites en détail dans le document *Normes de qualification harmonisées – Experts de contenus : besoins et exigences*. Les rédacteurs et réviseurs d'items doivent être titulaires d'un permis valide de représentant en assurance de personnes obtenu au Canada. Ils doivent aussi avoir une solide connaissance des produits, occuper un poste dans le cadre duquel ils assument des responsabilités liées à la vente ou à la formation, manifester un intérêt pour le processus de qualification et avoir une excellente aptitude à communiquer. Ils doivent aussi provenir de diverses régions géographiques et représenter différents types d'organisations : grandes entreprises, associations professionnelles et établissements de formation, entre autres. Les personnes bilingues sont encouragées à postuler. Pour que leur candidature à un poste de rédacteur ou de réviseur d'items soit considérée, les experts de contenu doivent soumettre aux OCRA un formulaire de candidature dûment rempli, une lettre d'intention et leur curriculum vitæ.

La *norme 4.8* précise que le concepteur d'un examen doit évaluer de nouveaux items à l'aide d'analyses empiriques ou en demandant à des experts de contenu d'en réviser le contenu. La formation donnée aux réviseurs doit être consignée, tout comme leurs qualifications et leurs caractéristiques sociodémographiques.

> ***Norme 4.8*** Le processus de révision du test devrait comprendre des analyses empiriques et, si nécessaire, devrait faire appel à des experts pour réviser les items et les critères de notation. Le cas échéant, les qualifications, l'expérience pertinente et les caractéristiques démographiques de ces experts devraient être documentées, tout comme les directives et la formation qu'ils reçoivent dans le cadre du processus de révision des items.

Une fois sélectionnés, les rédacteurs et les réviseurs des items de l'examen du PQAP prennent part à une formation sur la rédaction et la révision de questions. Le contenu de cette formation, qui est donnée par l'intermédiaire d'une présentation PowerPoint fournie par les OCRA, est conforme au document *Lignes directrices pour la rédaction et la validation : Questions d'examens du PQAP*. Les rédacteurs et les réviseurs d'items reçoivent des directives étape par étape sur la manière de rédiger ou de réviser des items; on leur transmet également des modèles appropriés.

Une fois qu'un item est rédigé, un spécialiste en mesure et évaluation le révise pour s'assurer de sa clarté, de sa pertinence par rapport aux spécifications de l'examen et de sa conformité aux règles de rédaction d'items. Le spécialiste communique ses commentaires au rédacteur de l'item et approuve les révisions. L'item est ensuite transmis à un réviseur qui le valide, puis le remet au spécialiste. Le spécialiste parachève la rédaction de l'item avant l'étape de la révision linguistique.

Pour assurer l'impartialité du processus et réduire les possibilités de biais, la révision est menée en aveugle : le réviseur ignore l'identité du rédacteur de l'item, et l'item qu'il révise est exempt d'information auxiliaire telle que la réponse exacte et la référence. Le réviseur a pour mandat d'évaluer l'item à la lumière des critères énoncés dans la grille d'analyse du contenu. Le réviseur doit en outre vérifier la réponse exacte, la référence de l'item et le lien vers l'élément de compétence figurant dans les spécifications de l'examen. Il doit aussi évaluer la pertinence de l'item pour la pratique professionnelle des agents en début de carrière, la fréquence d'utilisation du contenu dans la pratique et le degré de difficulté perçu de l'item. Le réviseur de l'item est censé suggérer des améliorations et apporter des changements à l'item si nécessaire.

Il est permis d'affirmer en conclusion que la sélection et la formation des experts de contenu sont bien documentées, mais que leurs caractéristiques ne le sont pas.

Conformité aux principes de rédaction d'items

Une évaluation psychométrique des items de l'examen du PQAP en version anglaise a été effectuée dans le cadre de la production de ce rapport. Tout d'abord, les items ont été évalués à la lumière des principes de rédaction d'items énoncés dans un article de Haladyna, Downing et Rodriguez publié dans une revue réputée (2002). Les problèmes relevés dans les items ont été notés, et une personne francophone faisant partie du personnel de Yardstick a vérifié si ces problèmes s'appliquaient aussi à la version française des items d'examen.

Les items de l'examen du PQAP sont, pour la plupart, d'une qualité jugée acceptable. Ils sont tous bien ciblés, clairs et basés sur des scénarios bien structurés. Quelques problèmes mineurs ont toutefois été décelés dans les items : certaines questions fermées nécessitent une réponse par « oui » ou « non », des prémisses contiennent le pronom « vous » et un manque de cohérence a été observé dans l'emploi des titres professionnels.

Dans l'examen du PQAP, quelques items nécessitent une réponse par « oui » ou « non ». La prémisse de ces items contient une question fermée à laquelle le candidat doit répondre par « oui » ou « non ». Toutefois, les choix de réponses ne se limitent pas à « oui » ou « non » : ils font aussi mention de la raison pour laquelle une mesure donnée doit être prise. La question n'amène pas le candidat à fournir une raison, mais celle-ci est énoncée dans les réponses. Les items appelant une réponse par « oui » ou « non » sont problématiques, parce que leur prémisse et leurs choix de réponses ne sont pas parfaitement en phase.

Dans d'autres items de l'examen du PQAP, la prémisse contient le pronom « vous ». Pour éviter que le candidat personnalise une question ou y réponde d'après son point de vue unique, il convient de formuler l'item à la troisième personne du singulier, en faisant référence à un agent, ou encore d'utiliser un titre professionnel.

Enfin, un manque de cohérence a été observé dans l'examen en ce qui a trait à l'emploi des titres professionnels. Par exemple, les titres de postes « représentant en assurance », « conseiller » et « spécialiste » sont utilisés de manière interchangeable. Il serait profitable de réviser l'ensemble des items de l'examen pour assurer l'uniformité de leur terminologie.

Équité de l'examen

Tous les programmes d'examen visent le même objectif crucial : assurer l'équité entre tous les candidats. L'équité peut se définir comme le fait de donner à tous les candidats une chance égale de faire valoir où ils se situent véritablement par rapport au construit mesuré par l'examen. Dans le contexte de l'élaboration d'items, l'équité fait référence à l'absence de contenu offensant, stéréotypé ou inhabituel susceptible de distraire les candidats et de les empêcher de mettre en valeur leur « vraie » compétence. Parmi les manquements à l'équité communément relevés dans des items figurent les suivants :

- Langage offensant;
- Contenu offensant;
- Contenu suscitant des réactions émotives;
- Contenu véhiculant des stéréotypes raciaux ou de genre;
- Références inégales aux hommes et aux femmes;
- Contenu peu familier à un groupe (p. ex., des acronymes ou des abréviations que certains groupes de candidats ne connaissent pas nécessairement);
- Vocabulaire peu familier à un groupe (p. ex., mots inusités ou complexes en français ou en anglais).

Quand il y a des raisons de croire que le contenu des items d'un examen (c.-à-d. des mots, des énoncés ou des phrases) pourrait être plus familier à certains groupes de candidats qu'à d'autres, il est important d'effectuer un examen de la sensibilité de l'ensemble des items de l'examen. Selon la *norme 3.2*, le concepteur d'un examen a la responsabilité de veiller à ce que le score ne soit pas influencé par les caractéristiques du candidat qui ne se rapportent pas à l'objectif de l'examen.

> ***Norme 3.2*** Les concepteurs de test ont le mandat d'élaborer des examens qui permettent de mesurer le construit voulu en réduisant au minimum le risque que des caractéristiques non pertinentes au contenu – qu'elles soient de nature linguistique, communicationnelle, cognitive, culturelle, physique ou autre – influencent le test.

Dans le cadre de l'élaboration des items, les OCRA ne procèdent pas à un examen de sensibilité formel. Cependant, des spécialistes des mesures et des évaluations ainsi que des réviseurs vérifient la clarté des items et la simplicité du langage employé. Selon les *normes*, « le niveau de maîtrise de la langue requis dans le test ne devrait pas excéder celui nécessaire au respect des exigences liées à l'exercice de la profession et à la certification, ou à la représentation du ou des construits visés » (p. 64). Autrement dit, le niveau de langue des items de l'examen du PQAP devrait être le même que celui employé par les

représentants en assurance de personnes en fonction. En plus de réviser les questions de l'examen pour veiller à ce qu'elles ne contiennent pas de termes complexes, les OCRA devraient se pencher sur les autres formes de contenu indélicat qu'elles pourraient receler.

Résumé et recommandations

- ***Sélection et formation des rédacteurs et des réviseurs d'items***. De façon générale, les processus fondamentaux relatifs à l'élaboration d'items sont suivis. Les OCRA fournissent aux rédacteurs et aux réviseurs d'items la formation et le soutien nécessaires pour les aider à élaborer des questions à choix multiple conformes aux spécifications de l'examen et aux principes généraux relatifs à la rédaction de ce type de questions. Les directives données aux rédacteurs et aux réviseurs d'items sont clairement documentées, à l'instar du processus général d'élaboration des items. De façon générale, les exigences de la *norme 4.8* sont respectées.

- ***Conformité aux principes de rédaction d'items.*** L'évaluation psychométrique des items français et anglais a révélé qu'ils sont rédigés conformément aux normes psychométriques (voir Haladyna *et al*., 2002). Les items sont de bonne qualité et ne nécessitent pas de révisions substantielles. Yardstick a cerné quelques problèmes mineurs dont les OCRA pourront tenir compte dans le cadre de l'élaboration de leurs directives de rédaction d'items. Ainsi, il n'est pas recommandé de formuler des questions auxquelles le candidat doit répondre par « oui » ou « non ». De même, le pronom personnel « vous » devrait être remplacé dans les items par un pronom à la troisième personne du singulier (c.-à-d. « il » ou « elle ») ou par un titre professionnel pertinent. Il pourrait être utile de réviser l'ensemble des items de l'examen pour assurer l'uniformité de leur terminologie.

- ***Équité de l'examen*.** Le concepteur d'un examen est tenu de faire en sorte que tous les candidats aient une chance égale de faire valoir leurs compétences dans le cadre de l'examen. Les items de l'examen du PQAP font actuellement l'objet d'une révision qui vise à assurer leur clarté et la simplicité du langage employé. En plus de réviser les questions de l'examen pour veiller à ce qu'elles ne contiennent pas de termes complexes, les OCRA devraient se pencher sur les autres formes de contenu indélicat qu'elles pourraient receler.

Étape III – Assemblage de l’examen

Dans le cadre de l’étape III, Yardstick s’est penchée sur la façon dont les formats d’examens ont été créés et a cherché à déterminer dans quelle mesure ils sont conformes aux paramètres fondamentaux de l’examen. Des analyses statistiques ont été réalisées pour évaluer la fiabilité de tous les formats d’examens et pour définir les propriétés psychométriques des items.

Assembler des questions pour créer un examen à des fins d’administration opérationnelle constitue une étape critique dans l’élaboration d’un examen. En documentant clairement le processus utilisé pour structurer la version finale de formats d’examens opérationnels ainsi que les éléments de preuve selon lesquels le contenu de l’examen concorde avec ses spécifications, on étaye la validité de l’interprétation des scores d’un examen.

Assemblage de l’examen

Dans le cadre de l’assemblage des formats d’examens, les OCRA utilisent des pondérations de compétence basées sur les spécifications de l’examen et des indices de difficulté théoriques des items. Ils s’assurent ainsi que la répartition des compétences des formats d’examens est appropriée et que ceux-ci ont un niveau de difficulté semblable.

La représentation des domaines des contenus joue un rôle incontournable dans la validation d’un examen. La *norme 4.12* précise d’ailleurs que les concepteurs d’un examen doivent mettre par écrit la méthode employée pour assembler la version finale de l’examen opérationnel. Il est notamment recommandé qu’ils fournissent une description de la manière dont le contenu de l’examen concorde avec les spécifications de l’examen. Ces renseignements permettront d’étayer la conclusion selon laquelle la performance des candidats à l’examen reflète leur compétence dans les domaines des contenus énoncés dans les spécifications de l’examen.

> ***Norme 4.12*** Les concepteurs de test devraient documenter dans quelle mesure le domaine des contenus d’un test représente le domaine tel que défini et les spécifications du test.

Les OCRA ont remis à Yardstick un document schématisant le lien entre les items de l’examen et les compétences issues des spécifications de l’examen. Comme on le constate à la lecture de ce document, les formats équivalents des quatre examens sont conformes aux spécifications de l’examen : cet élément constitue une preuve à l’appui de la *norme 4.12*.

Le document intitulé *Contrôle des examens du PQAP : Processus de transition et de maintenance* définit les critères d’assemblage d’un examen. L’assemblage d’un examen vise à répondre aux spécifications de l’examen et, parallèlement, à créer des formats équilibrés sur le plan statistique et en ce qui concerne les caractéristiques du contenu. Ce document établit une distinction entre les procédures d’assemblage d’examen en vigueur à l’heure actuelle (c.-à-d. qui s’appliquent à l’examen au cours de la période dite de transition) et celles qui le seront quand suffisamment de données auront été recueillies sur les items de l’examen (c.-à-d. au commencement de la période de maintenance).

À l’heure actuelle, les différents formats d’examens du PQAP contiennent environ le même nombre de mots. Ils renferment aussi une juste répartition d’items qui nécessitent des calculs, qui présentent une

étude de cas ou des choix de réponses plus longs ou qui ciblent des éléments précis d'une compétence. À l'avenir, il faudra également veiller à ce que les items des formats d'examens soient équilibrés en ce qui a trait au sexe, à la profession et aux caractéristiques sociodémographiques des personnes mentionnées dans les questions.

Au cours de la phase de transition de la mise en œuvre de l'examen, le niveau de difficulté des formats d'examens équivalents est pondéré à l'aide d'indices de difficulté théoriques des items. Un indice de difficulté théorique d'un item est une estimation du pourcentage de candidats répondant correctement à l'item en question. Les rédacteurs et les réviseurs d'items ont fourni de telles estimations pour toutes les questions faisant partie de la banque d'items de l'examen du PQAP.

Au cours de la période de transition, seuls les items ayant un indice de difficulté théorique de 40 % ou plus ou un indice de discrimination supérieur à 0,15 sont considérés pour l'assemblage des examens. L'indice de difficulté théorique des items est subjectif. Il ne s'agit pas d'un indicateur fiable de la qualité d'une question, contrairement à l'indice de difficulté statistique d'un item. Ce dernier repose sur les données d'un examen et, lorsqu'il est calculé sur de larges échantillons, constitue un indicateur fiable de la difficulté d'un item.

À l'heure actuelle, les OCRA se fondent sur des indices de difficulté théoriques des items pour concevoir de nouveaux formats d'examens et remplacer les questions qui, dans les formats actuels, ont une faible performance. Lorsqu'ils disposeront de plus de données sur l'examen du PQAP, les OCRA comptent utiliser des indices statistiques plutôt que des indices théoriques pour déterminer la qualité d'un item dans le cadre de l'assemblage de l'examen. Cette intention est exprimée dans le document *Contrôle des examens du PQAP : Processus de transition et de maintenance.*

À l'avenir, les critères déterminant la qualité d'un item deviendront légèrement plus stricts. Seuls les items ayant un indice de difficulté se situant entre 30 % et 85 % ou un indice de discrimination supérieur à 0,20 seront considérés pour l'assemblage des examens. En outre, les directives régissant l'assemblage des examens feront état de la répartition des items selon leur niveau de difficulté (faible : questions associées à une valeur de p de 30 à 49 % [10 % des items]; moyen : questions associées à une valeur de p de 50 à 69 % [60 % des items]; et élevé : questions associées à une valeur de p de 70 à 85 % [30 % des items]).

Malgré ces avancées positives, la sélection de nouveaux items pour l'examen demeurera fondée sur l'indice de difficulté théorique de la question. Pour remédier à ce problème, les OCRA auraient intérêt à procéder à un prétest des nouveaux items avant que ceux-ci apparaissent dans les examens opérationnels. De nouveaux items peuvent être ajoutés à titre expérimental aux examens opérationnels. Ils peuvent aussi faire l'objet d'un prétest auprès d'un échantillon de représentants en assurance de personnes ayant récemment obtenu leur permis.

Équivalence des formats d'examens

Les formats d'examens multiples servent souvent à contrer des problèmes de sécurité; la tricherie, par exemple. Quand des formats d'examens équivalents sont employés, il est primordial de s'assurer de l'équivalence des scores. L'équivalence des scores signifie qu'un candidat obtiendrait le même résultat à des examens dont le format est équivalent. Les scores associés à des formats d'examens équivalents sont interchangeables seulement quand les formats en question ont été conçus d'après les mêmes spécifications statistiques et relatives au contenu. Selon la *norme 5.12*, le concepteur d'un examen doit établir clairement, preuves à l'appui, dans quelle mesure des formats d'examens équivalents sont comparables au chapitre du contenu et des données statistiques.

> ***Norme 5.12*** Une justification claire et des éléments de preuve devraient être disponibles lorsqu'on soutient que les échelles de scores associées aux formats équivalents d'un test peuvent être utilisées de manière interchangeable.

Les OCRA satisfont à la *norme 5.12* en ce qui concerne les spécifications du contenu de l'examen. Des preuves attestent que les trois formats d'examens équivalents sont conformes aux spécifications de l'examen au chapitre de la représentation des compétences. Cela dit, aucun document ne semble contenir de données statistiques relatives à ces formats d'examens équivalents. Comme ils ne sont pas entièrement impartiaux, les indices de difficulté théoriques des items ne constituent pas une preuve irréfutable de l'équivalence des scores associés à des formats d'examens équivalents. Il est recommandé que les OCRA recueillent et consignent des renseignements sur la performance de formats d'examens équivalents dans le cadre de l'administration d'un même examen et de multiples examens : ils pourront ainsi affirmer avec certitude que les scores associés à ces formats d'examens sont équivalents.

Traduction et adaptation de l'examen

Quand un examen se décline en plusieurs langues, il est important que les scores associés à des formats d'examens multilingues aient la même signification. Selon les *normes*, la simple traduction d'un examen d'une langue à une autre ne garantit pas que le contenu et le niveau de difficulté de l'examen sont les mêmes que ceux de l'examen original. La traduction n'atteste pas non plus que la fiabilité et la validité de l'examen traduit sont semblables à celles de l'original. Les *normes* encouragent le concepteur d'un examen à examiner la validité, la fiabilité et la comparabilité des scores de formats d'examens qui se déclinent en différentes langues. Les *normes 3.12* et *7.6* précisent notamment que quand un examen est produit en deux langues ou plus, les méthodes de traduction et d'adaptation de l'examen devraient être décrites en détail. Il convient aussi de fournir des preuves de la fiabilité et de la validité des scores de l'examen.

> ***Norme 3.12*** Lorsqu'un test est traduit et adapté d'une langue à une autre, les concepteurs ou les utilisateurs du test sont tenus de décrire les méthodes employées pour établir l'équivalence de l'adaptation. Ils doivent également fournir des preuves empiriques ou logiques à l'appui de la validité de l'interprétation des scores du test selon les utilisations envisagées.

> ***Norme 7.6*** Quand un test est produit en deux langues ou plus, la documentation qui s'y rapporte doit faire état des méthodes employées pour traduire et adapter le test. Dans la mesure du possible, des renseignements doivent également être fournis sur les preuves à l'appui de la fiabilité/précision et de la validité du format adapté.

Les OCRA peuvent compter sur un processus rigoureux de traduction et d'adaptation des examens. Toutefois, le processus en question n'est pas bien documenté. Yardstick a obtenu des détails pertinents à ce sujet dans ses communications verbales avec les OCRA. Ces échanges ont permis d'établir clairement que les items d'un examen sont rédigés dans l'une des deux langues officielles, puis traduits

par les traducteurs de l'AMF. Par la suite, un expert de contenu révise les items à la lumière des spécifications de l'examen et les manuels de préparation à l'examen dans les deux langues. D'autres intervenants, les réviseurs linguistiques, collaborent à la traduction des items avec les rédacteurs, les traducteurs et les réviseurs. Toutes les parties concernées reçoivent des directives pour éviter d'employer, entre autres, des termes techniques qui ne sont pas expliqués dans les manuels de préparation à l'examen, des régionalismes, du jargon ou des unités de mesure propres à un contexte culturel particulier. Enfin, des spécialistes des mesures et des évaluations relisent les items en français et en anglais pour s'assurer de leur lisibilité et de leur clarté dans les deux langues. En conclusion, il est permis d'affirmer que les OCRA se fondent sur le jugement professionnel d'experts de contenu et de traducteurs pour attester l'équivalence des formats d'examens.

La *norme 3.12*, citée ci-dessus, précise en outre que le concepteur d'un examen devrait procéder à des analyses statistiques de formats d'examens en différentes langues pour vérifier si leur fiabilité est semblable. En principe, quand des formats d'examens sont équivalents, les caractéristiques des questions et de l'examen sont similaires. La valeur de p et le coefficient de corrélation bisériale de point corrigé des items doivent donc, eux aussi, être similaires. On ignore si les OCRA appliquent un processus de comparaison des statistiques des questions et de l'examen aux formats d'examens français et anglais. Un tel processus est recommandé. Si la taille des échantillons le permet, une analyse plus poussée du fonctionnement différentiel des items pourrait être réalisée à l'aide des méthodes de la théorie des réponses aux items.

Mise en équivalence des formats d'examens

Bien que les formats d'examens du PQAP soient élaborés selon les mêmes spécifications statistiques et relatives au contenu, leur niveau de difficulté réel peut différer : une mise en équivalence est alors nécessaire. Pour que les scores demeurent significatifs d'un examen à l'autre, l'examen du PQAP est mis en équivalence à l'aide d'un modèle d'examen à questions ancrées. En vertu de ce modèle, une batterie de questions communes (ancrées) est intégrée à tous les formats d'examens équivalents. Comme le stipule la *norme 5.15,* la batterie de questions ancrées est comme un mini-examen dont le contenu et le niveau de difficulté correspondent à ceux de la version intégrale de l'examen. Si aucune norme ne détermine le nombre exact de questions ancrées requises, il est recommandé d'élaborer une batterie de questions ancrées comportant de 20 à 30 items.

> ***Norme 5.15*** Dans les études d'équivalences utilisant le modèle de l'examen à questions ancrées, les caractéristiques de ce dernier et ses ressemblances avec les formats mis en équivalence devraient être présentées, en incluant à la fois les spécifications de contenu et les rapports empiriques entre les scores. Si des questions ancrées sont utilisées dans l'étude d'équivalences, la représentativité et les caractéristiques psychométriques des questions ancrées devraient être présentées.

Il y a 20 items dans l'*examen du module Déontologie – Common law* et 30 items dans chacun des trois autres examens modulaires, parmi lesquels figurent l'*examen du module Assurance-vie*, *l'examen du module Assurance contre la maladie ou les accidents* et l'*examen du module Fonds distincts*. Chaque format équivalent de l'examen du PQAP compte quatre questions ancrées, une pour chaque élément de compétence des spécifications de l'examen. La batterie de questions ancrées utilisée par les OCRA suscite deux observations. Premièrement, elle compte un nombre très restreint d'items. Bien que tous les éléments de compétence soient représentés dans la batterie de questions ancrées, ils ne semblent pas pondérés conformément aux spécifications de l'examen. En outre, il n'a pas été possible de déterminer

avec certitude si les OCRA ont tenu compte des propriétés psychométriques des questions dans le cadre de la création de la batterie de questions ancrées. Pour assurer la qualité d'une mise en équivalence, la batterie de questions ancrées doit refléter étroitement le contenu et les propriétés psychométriques de la version intégrale de l'examen. Par conséquent, il serait profitable pour les OCRA d'accroître le nombre d'items de sa batterie de questions ancrées et de sélectionner celles-ci sur la base des caractéristiques des items. Yardstick est consciente du fait que les formats d'examens du PQAP sont courts et que, par conséquent, il pourrait être impossible d'ajouter des items à la batterie de questions ancrées. Le choix d'une autre méthode de mise en équivalence pourrait constituer une solution de rechange.

Le processus de comparaison de la performance des formats d'examens équivalents, une fois les données sur l'examen recueillies, n'est étayé par aucun document. On peut présumer que les OCRA évaluent la performance statistique de la batterie de questions ancrées dans l'ensemble des formats d'examens. Or, selon la *norme 5.17*, des preuves directes de la comparabilité des scores de formats d'examens équivalents doivent être produites.

> ***Norme 5.17*** Quand un lien est établi entre des scores obtenus à des tests qui ne peuvent être mis en équivalence, il convient de fournir des preuves directes de la comparabilité de ces scores. De même, la population de candidats auxquels s'applique la comparabilité des scores doit être clairement mentionnée. La justification et les preuves requises dépendent en partie des utilisations envisagées pour lesquelles la comparabilité des scores s'avère nécessaire.

Résumé et recommandations

- ***Assemblage de l'examen.*** Conformément aux *normes*, l'examen du PQAP est assemblé sur la base des pondérations de la compétence issues des spécifications de l'examen. La correspondance entre le contenu et les spécifications de l'examen est bien documentée.

  Dans le cadre de l'assemblage de formats d'examens équivalents, les OCRA équilibrent le niveau de difficulté de ces formats au moyen des indices de difficulté théoriques des items qui leur ont été fournis par des experts de contenu. Comme ces indices sont intrinsèquement subjectifs, ils peuvent entraîner une sous-estimation ou une surestimation du niveau de difficulté réel des items. Les indices statistiques de difficulté des items constituent un indicateur beaucoup plus fiable de la qualité d'un item, car ils reposent sur des données objectives issues d'un vaste échantillon de candidats.

  Il serait profitable que les OCRA procèdent à un prétest des nouvelles questions pour obtenir des indices statistiques de difficulté et de discrimination des items et qu'ils se servent de ces deux indices pour assembler l'examen. Ce prétest pourrait être effectué en intégrant les nouveaux items aux formats d'examens opérationnels ou par l'intermédiaire d'un prétest formel auquel se prêteraient des représentants en assurance de personnes ayant obtenu leur permis depuis peu.

- ***Équivalence des formats d'examens.*** Des formats d'examens sont jugés équivalents quand ils sont conçus d'après les mêmes spécifications statistiques et relatives au contenu. Les OCRA disposent de preuves à l'appui du fait que les formats équivalents des examens modulaires du PQAP ont un contenu semblable. Toutefois, l'équivalence de ces formats n'est étayée par

aucune donnée statistique documentée. Il est recommandé que les OCRA recueillent et consignent des renseignements sur la performance de formats d'examens équivalents dans le cadre de l'administration d'un même examen et de multiples examens. L'étape V du présent rapport contient plus de renseignements sur l'équivalence des formats d'examens : elle renferme notamment une description des résultats des analyses statistiques réalisées par Yardstick d'après des formats d'examens équivalents.

- ***Traduction et adaptation de l'examen.*** Comme l'examen du PQAP se décline dans les deux langues officielles, il est important d'établir que les scores associés aux versions française et anglaise des examens ont la même signification. Les OCRA se fondent sur le jugement professionnel de plusieurs experts de contenu et de traducteurs pour s'assurer que le contenu de leurs formats d'examens est équivalent dans les deux langues. Par ailleurs, les OCRA devraient procéder à une recherche empirique sur les versions française et anglaise des examens pour déterminer si celles-ci sont semblables au chapitre de la fiabilité de l'examen et des statistiques sur les items. Par exemple, les statistiques sur les items français et anglais pourraient être comparées. Une analyse du fonctionnement différentiel des items pourrait aussi être menée. L'examen du contenu et les recherches empiriques sur les formats d'examens français et anglais devraient être documentés en détail : les OCRA auront ainsi une preuve tangible de l'équivalence des formats.

- ***Mise en équivalence des formats d'examens.*** Les exigences de mise en équivalence des formats d'examens énoncées dans les *normes* sont partiellement respectées. Les OCRA utilisent un modèle d'examen contenant une batterie de questions ancrées pour mettre en équivalence les différents formats des examens modulaires du PQAP. Idéalement, une batterie de questions ancrées doit refléter le contenu et les propriétés psychométriques de la version intégrale d'un examen. Bien que les formats d'examens du PQAP comportent des questions ancrées applicables à chaque élément d'une compétence, le nombre de questions ancrées est trop restreint pour qu'une comparaison significative puisse être établie entre les formats d'examens. Pour assurer la qualité de la mise en équivalence d'un examen, la batterie de questions ancrées doit refléter étroitement le contenu et les propriétés psychométriques de l'ensemble de l'examen. Dans cette optique, il serait préférable que chaque format d'examen des OCRA soit constitué de questions ancrées dans une proportion d'au moins 25 % et que ces questions constituent un mini-examen dont le contenu, le niveau de difficulté et l'indice de discrimination correspondent à ceux de la version intégrale de l'examen.

  Si le nombre de questions ancrées dans l'examen augmente de façon significative, les différents formats de l'examen se chevaucheront davantage. Si un tel chevauchement n'est pas souhaitable, une autre méthode de mise en équivalence devrait être envisagée. Par exemple, des formats d'examens pourraient être mis en équivalence sur la base du seuil de réussite des items déterminé dans le cadre de l'établissement de la norme. L'établissement de la norme est décrit en détail à l'étape VI du présent rapport. Contrairement aux statistiques sur les items, le seuil de réussite des items ne varie pas en fonction de l'échantillon : il s'agit donc d'une bonne façon de calibrer les items selon leur niveau de difficulté. Le seuil de réussite des items peut être utilisé au cours de l'assemblage de l'examen pour concevoir des formats d'examens équivalents associés à la même norme de réussite.

Étape IV – Administration, notation et communication des scores de l'examen

Dans le cadre de l'étape IV, Yardstick s'est attardée aux conditions d'administration de l'examen du PQAP : elle a notamment pris en compte les directives données aux candidats, les responsabilités du personnel responsable de l'administration de l'examen, les procédures de sécurité liées à la passation de l'examen, l'environnement dans lequel les candidats passent l'examen et le processus d'administration de l'examen. Les méthodes de notation et de communication des scores ont aussi été évaluées.

L'exactitude et la pertinence des scores d'un examen sont tributaires de la standardisation des méthodes d'administration, de notation et de communication des scores de l'examen. La standardisation de ces méthodes permet de faire en sorte que tous les candidats aient des chances égales de faire valoir leurs compétences sans qu'aucun ne bénéficie d'un avantage déloyal. Dès que les méthodes et processus d'administration, de notation et de communication des scores d'un examen ne sont pas uniformes, la validité de l'interprétation des scores est compromise.

Administration de l'examen

La *norme 6.1* recommande que le concepteur d'un examen mette au point des procédures d'administration et de notation de l'examen, fournisse de la formation et du soutien aux personnes responsables de la mise en œuvre de ces procédures et veille à l'application de ces procédures en exerçant une supervision. Par exemple, les directives données aux candidats et le délai imparti pour la passation de l'examen doivent être précisés et respectés à la lettre. Les surveillants d'examen doivent recevoir une formation sur la manière de créer des conditions d'administration d'examen standardisées qui favorisent l'interprétation des scores selon les utilisations envisagées. Ils doivent être conscients de leur rôle et de leurs responsabilités et savoir quels renseignements transmettre aux candidats avant, pendant et après l'examen. Des directives doivent en outre faire état des situations dans lesquelles il est permis de déroger aux conditions standardisées d'administration d'un examen pour les candidats qui nécessitent des accommodements particuliers. Enfin, le concepteur de l'examen doit élaborer une politique relative à la reprise de l'examen.

> ***Norme 6.1*** Les personnes qui font passer un test doivent se conformer attentivement aux procédures standardisées de passation et de notation établies par le concepteur du test et aux directives énoncées par l'utilisateur du test.

Pour l'heure, l'administration de l'examen du PQAP relève des juridictions participantes. Le concepteur de l'examen remet aux juridictions le *Practical Exam Administration Guidelines*, un document qui énonce les paramètres généraux d'administration, de notation et de communication des scores d'un examen. Dépendamment de la politique, le document *Practical Exam Administration Guidelines* peut fournir des suggestions judicieuses aux juridictions (p. ex., politique sur la reprise de l'examen) ou leur donner une autonomie décisionnelle considérable (p. ex., politique sur les accommodements particuliers).

De façon générale, les juridictions participantes semblent avoir beaucoup de latitude dans leur manière de faire passer l'examen du PQAP. Par exemple, les juridictions déterminent l'admissibilité des candidats à l'examen du PQAP et sont autorisées à accorder des exemptions. Elles décident aussi du

mode d'administration de l'examen (par ordinateur ou sur papier) et définissent les procédures d'identification des candidats et celles ayant trait aux accommodements particuliers.

Le fait que les juridictions puissent agir sur des composantes importantes de l'administration de l'examen ouvre la porte à des incohérences dans les conditions d'administration de l'examen. Autrement dit, les candidats de différentes juridictions pourraient ne pas passer l'examen dans les mêmes conditions. Des différences dans les conditions d'administration d'un examen peuvent avoir une incidence négative sur les résultats obtenus par un candidat, en plus de porter préjudice à la validité des scores et à l'équité de l'examen.

Par exemple, dans une juridiction, les candidats peuvent devoir passer tous les examens modulaires le même jour. En revanche, dans une autre juridiction, ils peuvent ne devoir passer qu'un examen modulaire par jour. Comme les candidats de la première juridiction sont probablement plus fatigués au moment d'entreprendre leur dernier examen modulaire, les résultats qu'ils obtiennent à cet examen ne seront pas nécessairement à la hauteur de leurs compétences. Les conditions de passation de l'examen dans cette juridiction peuvent donc empêcher les candidats de mettre en valeur leurs véritables connaissances et d'obtenir le score escompté. Par opposition, les conditions de passation de l'examen dans l'autre juridiction n'auront pas une telle incidence sur les scores des candidats.

La standardisation des politiques et procédures de l'examen aidera les OCRA à éliminer la variance non pertinente au contenu dans les scores obtenus. La variance non pertinente au contenu renvoie à une variance dans les scores des candidats qui ne peut être expliquée par la mesure du contenu de l'examen. Des procédures d'accommodements particuliers qui diffèrent d'une juridiction à l'autre, par exemple, constituent une source potentielle de variance non pertinente au contenu dans les scores d'un examen : elles devraient donc être évitées dans le cadre de l'administration d'un examen.

La *norme 6.4* précise qu'un examen doit se dérouler dans un environnement confortable avec un minimum de distractions. La *norme 6.5* stipule quant à elle que les candidats doivent recevoir les renseignements nécessaires sur l'examen, y compris des documents contenant des exercices. Ainsi, non seulement les conditions d'examen doivent être standardisées d'une juridiction à l'autre, mais les candidats doivent aussi être informés de ces conditions avant l'examen.

> ***Norme 6.4*** Le lieu utilisé pour la passation d'un test devrait offrir aux candidats un confort raisonnable et comporter le moins de distractions possible pour éviter la variance non pertinente au contenu.

> ***Norme 6.5*** Des instructions et des exercices appropriés ainsi que d'autres mesures de soutien devraient être proposés aux candidats pour atténuer la variance non pertinente au contenu.

La *norme 8.1* précise qu'un organisme doit fournir aux candidats des renseignements de base sur l'objectif et le contenu d'un examen ainsi que sur le processus d'administration de l'examen pour que l'accès à l'information soit équitable parmi les candidats. L'organisme doit informer les candidats du contenu abordé dans l'examen, notamment du domaine de compétence qui sera évalué et du format des items, pour les aider à se préparer à l'examen. Selon la *norme 8.2*, l'organisation doit aussi

communiquer aux candidats des renseignements sur la procédure d'administration d'un examen, les critères de notation, l'utilisation envisagée des scores de l'examen, l'accès à des accommodements particuliers, la politique de reprise de l'examen et la protection de la confidentialité. Bon nombre d'organisations choisissent d'intégrer ces renseignements à un manuel du candidat accessible sur leur site Web. Ce document, qui fournit au candidat de l'information logistique fort utile, vise à atténuer la variance non pertinente au contenu dans le cadre de l'administration de l'examen.

> ***Norme 8.1*** Les renseignements relatifs au contenu et aux objectifs d'un test préparés à l'intention des candidats avant l'administration du test devraient être offerts à tous. Les renseignements devraient être communiqués gratuitement et selon un format facilement accessible.

> ***Norme 8.2*** Le cas échéant, les candidats devraient recevoir à l'avance tous les renseignements possibles concernant le test, le processus d'administration du test, les utilisations prévues, les critères de notation, la politique d'administration du test, l'accès à des accommodements et la protection de la confidentialité, tout en assurant la validité des réponses et l'interprétation adéquate des scores du test.

La *norme 6.2* porte plus particulièrement sur les procédures d'accommodement prévues dans le cadre de l'examen. Les candidats doivent connaître les procédures d'accommodement auxquelles ils ont accès et le processus qu'ils doivent suivre pour en bénéficier.

> ***Norme 6.2*** Lorsque des procédures formelles ont été établies pour requérir un accommodement et en bénéficier, les candidats doivent en être informés avant la passation du test.

D'après les *Guidelines for the Implementation of the LLQP*, les juridictions ont la responsabilité d'adapter le matériel d'évaluation pour les candidats en vertu des principes généraux d'équité énoncés dans le document. Il convient de noter que seul le concepteur de l'examen a l'expertise nécessaire pour déterminer les modifications qu'il convient d'apporter à l'examen, à la documentation qui s'y rattache ou à ses conditions d'administration. Ces modifications doivent être effectuées d'une manière conforme à l'objectif de l'examen. Autrement dit, l'examen doit permettre de mesurer un construit déterminé chez tous les candidats, y compris ceux qui ont réclamé des accommodements.

Les *normes 6.6 et 6.7* exigent que les organismes déploient des efforts raisonnables pour assurer l'intégrité des scores d'un examen et pour protéger en tout temps la sécurité du matériel d'administration de l'examen.

Norme 6.6 Des efforts raisonnables devraient être consentis pour assurer l'intégrité des scores d'un test en éliminant les possibilités de recourir à des moyens frauduleux ou trompeurs pour obtenir un résultat.

Norme 6.7 Les utilisateurs de test ont la responsabilité de protéger en tout temps la sécurité du matériel d'administration du test.

À l'heure actuelle, le concepteur de l'examen laisse aux juridictions le soin d'établir des procédures appropriées d'identification des candidats et d'aménager la salle d'examen. Le document *Guidelines for the Implementation of the LLQP* recommande que les juridictions confirment l'identité du candidat et « fassent en sorte que l'examen se déroule dans des conditions propices à la réussite » (p. 6). Il mentionne aussi que « les candidats doivent être suffisamment à l'aise, de manière à ce qu'aucun facteur externe ne fasse entrave à leurs résultats » (p. 6). Ces recommandations sont relativement générales, au sens où elles ne précisent pas comment les salles d'examen doivent être aménagées au juste pour réduire au minimum les risques de tricherie.

Les juridictions sont aussi tenues de veiller à l'intégrité du matériel d'administration de l'examen. Selon les *Guidelines for the Implementation of the LLQP*, ce matériel et les papiers brouillons doivent être ramassés à la fin de la période d'examen. Le document stipule aussi qu'un surveillant doit superviser le déroulement de l'examen, et que toute irrégularité survenant pendant l'examen doit être portée à l'attention des autorités compétentes à l'échelle de la juridiction. Cela dit, il n'est pas clairement établi si des mesures sont en place pour protéger la sécurité du matériel d'administration de l'examen avant et après l'examen. Les directives données aux surveillants ne sont pas précisées non plus, ni les procédures engagées quand un candidat est soupçonné de tricherie.

Notation de l'examen

La qualité de la notation d'un examen a une incidence directe sur la fiabilité et la validité de l'interprétation des scores de l'examen. La notation doit être exacte et cohérente d'un candidat à l'autre, car elle doit constituer une assise solide aux décisions prises au sujet des candidats en fonction des scores qu'ils ont obtenus. Le risque que des erreurs se glissent dans le processus de notation est réduit au minimum, voire éliminé, si des processus d'assurance de la qualité sont en place (p. ex., vérification de la clé de réponses, double notation par un tiers ou vérification manuelle des scores obtenus à l'examen par un échantillon de candidats sélectionnés au hasard). En fait, selon les *normes 6.8 et 6.9*, un organisme doit établir un protocole de notation et veiller à ce que la notation soit fondée sur des processus de contrôle de la qualité dûment documentés.

Norme 6.8 Les personnes responsables de la notation d'un test devraient établir des protocoles de notation. Lorsque la notation implique que des personnes aient à porter des jugements, ces personnes devraient se fonder sur des rubriques, des procédures et des critères de correction. Quand la notation de réponses complexes est générée par ordinateur, l'exactitude de l'algorithme et des processus devrait être documentée.

> ***Norme 6.9*** Les personnes responsables de la notation d'un test devraient établir des processus et des critères de contrôle de la qualité et veiller à ce que ceux-ci soient bien documentés. Une formation adéquate devrait être fournie. La qualité de la notation devrait être surveillée et consignée. Toute source systématique d'erreurs de notation doit être documentée et corrigée.

La notation de l'examen du PQAP relève des juridictions participantes. Selon les *Guidelines for the Implementation of the LLQP*, « chaque juridiction est responsable de noter les différentes versions d'un examen à l'aide des outils de son choix, en se basant sur la clé de réponses appropriée qui lui a été fournie » (p. 9). Les procédures d'assurance de la qualité mises en place pour valider l'exactitude de la notation de l'examen ne sont pas clairement définies. Comme il est possible de passer l'examen du PQAP par ordinateur ou sur papier, les procédures de contrôle de la qualité et de notation applicables aux deux modes d'administration peuvent varier d'une juridiction à l'autre.

Communication des scores de l'examen

Lorsque l'on communique les résultats d'un examen aux candidats, il faut veiller à ce que ces résultats soient interprétés correctement par les principaux intéressés. Il s'agit dans le cas présent des candidats eux-mêmes et des autorités responsables de la délivrance des permis, lesquelles prennent des décisions à l'égard des candidats à la lumière des scores obtenus à l'examen. Les candidats doivent être informés de leur degré de performance à l'examen et de la manière dont leur performance est évaluée en vue de l'obtention de leur certification. Les autorités responsables de la délivrance des permis doivent se servir des scores d'un examen uniquement pour prendre des décisions relatives à l'octroi de permis. Par exemple, ces autorités ne doivent pas permettre aux employeurs d'utiliser les scores d'un examen à des fins de recrutement ou de promotion. Les scores d'un examen ne doivent être utilisés qu'aux fins pour lesquelles l'examen a été élaboré.

Selon les *normes*, les organismes doivent expliquer quelles sont les fins envisagées et non envisagées de l'information contenue dans le relevé du score. D'après la *norme 6.10*, la communication des scores doit comprendre une description de la portée de l'examen (p. ex., les principaux domaines de compétences), de ce que les scores représentent (p. ex., compétence versus manque de compétence), de la fiabilité des scores et de la façon dont les scores devraient ou ne devraient pas être utilisés.

> ***Norme 6.10*** Lorsque les scores d'un test sont communiqués, les personnes responsables du programme d'administration du test doivent fournir les interprétations appropriées aux personnes concernées. Ces interprétations doivent décrire en langage simple la portée du test, la signification, la précision et la fiabilité des scores ainsi que l'utilisation prévue des scores.

La responsabilité de la communication des résultats de l'examen du PQAP relève des juridictions. En ce moment, chaque juridiction détermine elle-même quels types de renseignements elle communique aux candidats en ce qui concerne les résultats de l'examen. En outre, il appartient à chaque juridiction de choisir le moment où elle communique ces résultats et la manière dont elle s'y prend. Il est recommandé de communiquer les résultats de l'examen aux candidats d'une manière structurée. Le

relevé du score doit au moins faire état du nombre de questions d'examen par domaine de compétences, du résultat obtenu par le candidat (p. ex., réussite ou échec) et du score total de l'examen.

En vertu des *normes*, un organisme doit avoir une politique régissant la conservation des examens et leur utilisation potentielle. Les *Guidelines for the Implementation of the LLQP* précisent que « le support sur lequel les candidats transcrivent leurs réponses » doit être conservé pendant cinq ans. Aucun autre renseignement n'est fourni sur les procédures de conservation de l'information relative à l'examen.

> ***Norme 6.14*** Les organismes conservant de l'information non anonyme sur les scores d'un test devraient élaborer une politique visant à encadrer clairement la durée de conservation des dossiers individuels, l'accès aux données contenues dans ces dossiers et l'utilisation de ces données (à des fins de recherche ou autre) au fil du temps. La politique devrait être documentée et accessible aux candidats. Il appartient aux utilisateurs du test d'assurer la sécurité des données sur les plans administratif, technique et physique.

Résumé et recommandations

- ***Procédures d'administration de l'examen.*** Les scores d'un examen sont plus susceptibles d'avoir la même signification et d'être facilement interprétables quand les conditions d'administration de l'examen sont standardisées. Yardstick recommande aux OCRA de régulariser toutes les composantes du processus d'administration de leur examen en créant un ensemble de politiques et procédures exhaustives auxquelles toutes les juridictions doivent se conformer lorsqu'elles font passer l'examen aux candidats.

- ***Accommodements consentis aux candidats à l'examen.*** Yardstick recommande aux OCRA de dresser la liste des accommodements dont les candidats peuvent se prévaloir et de définir clairement les critères à respecter pour demander et accorder ces accommodements. Par ailleurs, les procédures de demande d'accommodements dans le cadre de l'examen doivent être communiquées explicitement aux candidats. La création d'une politique détaillée sur les accommodements consentis aux candidats à l'examen dans toutes les juridictions et la communication de cette politique aux candidats contribuera à assurer l'équité de l'examen et la validité de l'interprétation des scores.

- ***Intégrité de l'examen.*** Yardstick recommande au concepteur de l'examen de créer un ensemble de politiques et procédures détaillées pour encadrer l'administration de l'examen, lesquelles devront être mises en œuvre par toutes les juridictions. Ce document permettra de standardiser les processus inhérents à l'administration de l'examen dans toutes les provinces, en plus de favoriser l'équité de l'examen et la validité de l'interprétation des scores. Le document devra décrire en détail les procédures à suivre avant, pendant et après l'administration de l'examen du PQAP. Il peut notamment contenir les renseignements suivants :

    - Directives à appliquer avant l'examen
    - Directives à appliquer le jour de l'examen (y compris les directives verbales à donner aux candidats)

- Procédures à suivre pendant la passation de l'examen
- Procédures à suivre à la fin de la passation de l'examen
- Directives à appliquer après l'examen

Idéalement, ce document doit faire état des critères d'identification des candidats. Il ne devrait pas appartenir aux juridictions de déterminer ce qu'est une méthode acceptable de vérification de l'identité des candidats. Ce document devra aussi contenir des directives à l'intention des surveillants, lesquelles visent à assurer la manipulation sécuritaire du matériel d'administration de l'examen avant, pendant et après l'examen, et décrire le protocole à suivre pour signaler des irrégularités.

- ***Notation de l'examen.*** Il serait profitable que les OCRA mettent par écrit des procédures appropriées de notation et d'assurance de la qualité et qu'ils les communiquent aux juridictions. Les processus de notation et d'assurance de la qualité doivent être synchronisés entre les juridictions. La meilleure façon de réussir à standardiser la notation d'un examen est de la centraliser. En outre, le concepteur de l'examen peut décider d'établir un processus de double notation ou de vérification manuelle des réponses pour un sous-groupe de candidats.

- ***Communication des scores de l'examen.*** Pour assurer la validité de l'interprétation des scores d'un examen, les juridictions doivent communiquer aux candidats les mêmes renseignements sur la performance; cette communication doit au demeurant se faire de la même façon d'une juridiction à l'autre. Autrement dit, la standardisation de la communication des scores d'un examen s'impose.

  Le relevé du score doit au moins faire état du contenu abordé dans l'examen, du résultat obtenu (p. ex., réussite ou échec) ainsi que du score total obtenu à l'examen ou du score obtenu à chaque examen modulaire. Il est à noter que certains organismes de réglementation fournissent de l'information supplémentaire aux candidats qui échouent à l'examen en précisant, dans le relevé du score, le résultat obtenu à chaque composante de l'examen. Ce mode de communication des scores ne convient pas à l'examen du PQAP, chacune de ses composantes contenant un petit nombre d'items. Moins il y a d'items pris en compte dans le calcul d'un score, moins le score a de chances d'être fiable. De façon générale, le relevé du score doit rendre compte du score total obtenu à l'examen. Ce score est en effet plus susceptible d'être fiable que ceux obtenus à chacune des composantes de l'examen, car il repose sur un plus grand nombre d'items.

Étape V – Analyse d'items et d'examen

Dans le cadre de l'étape V, Yardstick a examiné les méthodes statistiques employées par les OCRA pour évaluer la qualité générale de l'examen du PQAP et celle de ses items. Yardstick a aussi procédé à des analyses d'items et d'examen sur tous les formats des quatre examens modulaires, de manière à vérifier de façon indépendante la qualité de l'examen et des questions.

Les méthodes d'évaluation statistique de l'examen du PQAP sont décrites dans les documents suivants : *Contrôle des examens du PQAP : Processus de transition et de maintenance, Processus de renouvellement des questions d'examens* et *Rapport sur les mesures transitoires : Implantation du PQAP*. Cette documentation tend à prouver la conformité à la *norme 4.7*, laquelle encourage les concepteurs d'un examen à consigner par écrit les méthodes de sélection et de test appliquées aux items.

> ***Norme 4.7*** Les méthodes utilisées pour élaborer, réviser et expérimenter les items et les choisir à partir d'un regroupement d'items devraient être documentées.

Les OCRA évaluent la fiabilité et la variabilité des scores d'un examen, le niveau de difficulté des items et la capacité des items à établir une distinction entre les candidats forts et les candidats faibles.

La fiabilité de l'examen est évaluée à l'aide de l'alpha de Cronbach, un indice statistique qui reflète la cohésion de questions servant à mesurer un construit unique et unidimensionnel. L'alpha de Cronbach se situe entre 0,0 et 1,0, la valeur étant directement proportionnelle au degré de fiabilité de l'examen. De façon générale, un coefficient alpha de Cronbach supérieur à 0,70 est jugé acceptable. Cela dit, dans le cas d'examens certificatifs aux enjeux élevés, il est préférable de viser un alpha de Cronbach d'au moins 0,80.

La valeur de p et le coefficient de corrélation bisériale de point corrigé d'un item rendent compte de son niveau de difficulté et de son pouvoir de discrimination. La valeur de p est une mesure de difficulté d'un item : elle représente la proportion de candidats répondant correctement à l'item en question. Cette valeur se situe entre 0 et 1,0 : plus elle est faible, plus le niveau de difficulté de l'item est élevé. En général, les items dont la valeur de p est inférieure à 0,30 sont jugés très difficiles par les candidats, tandis que ceux sont la valeur de p est supérieure à 0,90 ou à 0,95 sont jugés très faciles.

Le coefficient de corrélation bisériale de point corrigé est un rapport entre le score obtenu à un item et le score total obtenu à l'examen après soustraction du score obtenu à l'item en question. En se basant sur ce facteur, il est possible de déterminer dans quelle mesure un item permet de distinguer les candidats forts des candidats faibles. Le coefficient de corrélation bisériale de point corrigé se situe entre -1,0 et 1,0. Une valeur positive élevée signifie que l'item contribue de façon significative au score total de l'examen : les candidats qui ont répondu correctement à l'item ont également obtenu un bon résultat à l'examen. Les items dont le coefficient de corrélation bisériale de point corrigé est supérieur à 0,15 sont jugés acceptables.

À l'heure actuelle, les OCRA évaluent la performance statistique des items de leurs examens sur une base hebdomadaire. Selon le document *Contrôle des examens du PQAP : Processus de transition et de maintenance*, quand le taux de réussite à un examen est inférieur à 70 %, les scores obtenus par certains candidats sont ajustés : un ou deux items ayant une faible valeur de p sont alors retirés. Après

l'administration de 300 examens, les formats d'examens sont réévalués et révisés pour veiller à ce que la valeur de p de leurs items soit supérieure à 40 % et l'indice de discrimination, supérieur à 0,15.

La *norme 4.10* stipule que le concepteur d'un examen doit se fonder sur un échantillon de taille adéquate pour procéder à l'évaluation psychométrique des items, expliquer le processus par lequel les items sont choisis et documenter les résultats des analyses statistiques.

> ***Norme 4.10*** Lorsque le concepteur d'un test évalue les propriétés psychométriques des items, le modèle utilisé à cette fin (p. ex., la théorie classique des tests, la théorie des réponses aux items ou un autre modèle) devrait être documenté. L'échantillon utilisé pour estimer les propriétés des items devrait être décrit. Il devrait aussi être d'une diversité et d'une taille adéquates pour ce genre d'analyse. Le processus par lequel les items sont choisis et les données utilisées pour le choix d'items, notamment la difficulté des items, leur discrimination ou le fonctionnement différentiel des items pour les grands groupes de candidats, devrait également être documenté.

En règle générale, les pratiques des OCRA relatives à l'évaluation psychométrique des examens sont conformes à la *norme 4.10*. Le fait que les OCRA se basent sur des données issues de vastes échantillons de candidats pour réévaluer et réviser leurs formats d'examens constitue une pratique digne de mention. De même, les critères statistiques d'évaluation des items sont décrits de A à Z, par écrit.

Cela dit, l'ajustement sélectif des scores obtenus à un examen par les candidats faisant partie d'une cohorte faible soulève certaines préoccupations. Il n'est pas approprié de réévaluer les scores d'un examen seulement quand le taux de réussite est faible, car une telle pratique porte ombrage aux scores des candidats qui ont échoué à l'examen alors que le taux de réussite était élevé. Pourquoi les résultats de ces candidats n'ont-ils pas été ajustés par le retrait des items ayant une piètre performance?

Quand commencera la période de maintenance de l'examen du PQAP, les OCRA appliqueront de nouvelles méthodes d'évaluation statistique des items. Ces nouvelles méthodes, qui sont énoncées dans le document *Contrôle des examens du PQAP : Processus de transition et de maintenance*, consistent notamment en des analyses statistiques mensuelles de formats d'examens qui visent à cibler les items ayant une piètre performance pour les remplacer par d'autres dont la performance est meilleure. Les items ayant une piètre performance sont ceux dont la valeur de p est inférieure à 30 % ou supérieure à 85 % et ceux dont l'indice de discrimination est plus bas que 0,20.

Les formats d'examens seront mis à jour environ deux fois l'an. Dans le cadre de la sélection d'items pour de nouveaux examens, les OCRA accorderont la priorité aux questions dont l'indice de discrimination est élevé : cette stratégie vise à faire en sorte que la fiabilité des formats d'examens atteigne 0,70. Les nouvelles méthodes d'évaluation statistique de l'examen du PQAP sont raisonnables.

Il sera profitable pour les OCRA de consigner les résultats de toutes les analyses d'items et d'examen qui seront réalisées à l'avenir pour l'examen du PQAP. Faire le suivi de la performance statistique de multiples formats d'examens au fil du temps aidera les OCRA à recueillir des éléments de preuve à l'appui de la validité de l'interprétation des scores de l'examen. Par exemple, en vertu de la *norme 2.3*, le concepteur d'un examen doit estimer la fiabilité de l'examen et en faire mention dans le relevé du

score. Il est clair que les OCRA disposent des renseignements pertinents à l'interne. Ces renseignements doivent cependant être compilés après l'administration de chaque examen.

> ***Norme 2.3*** Pour chaque score total, sous-score ou combinaison de scores devant faire l'objet d'une interprétation, une estimation des indices de fiabilité ou de précision pertinents doit être communiquée.

En vue de procéder à une évaluation statistique indépendante de l'examen du PQAP, Yardstick a réalisé des analyses d'items et d'examen pour 12 formats des quatre examens modulaires soumis aux candidats au printemps 2016. L'examen du PQAP est employé depuis janvier 2016. Entre les mois de janvier et mai 2016, les formats d'examens ont été mis à jour deux fois pour améliorer la qualité des items. Dans le cadre de ces mises à jour, des items ont été remplacés à la lumière des résultats d'analyses statistiques.

Chaque fois que le format d'un examen était mis à jour, une nouvelle version du format en question était créée. Le tableau 1 présente les différentes versions des formats d'examens du PQAP, la période d'utilisation des versions en question et le nombre moyen de candidats par format d'examen. Il indique aussi quelles versions de l'examen Yardstick a sélectionnées pour ses analyses d'items et d'examen. Pour assurer l'intégrité des résultats, une analyse d'items et d'examen a été réalisée pour une seule version d'un examen. Il n'a pas été possible de compiler les données relatives à l'examen dans les différentes versions.

Tableau 1. Données sur les examens des modules Déontologie – Common law, Assurance-vie, Assurance contre la maladie ou les accidents et Fonds distincts (versions anglaises seulement)

Version de l'examen	Période d'utilisation opérationnelle	Nombre moyen de candidats par format d'examen	Inclus dans les analyses par Yardstick? Oui/Non
1, 2 et 3	Du 4 janvier au 17 mars 2016 (Ontario) Du 7 janvier au 7 mars 2016 (toutes les autres provinces)	374	Non
4.1, 5.1 et 6.1*	Du 18 mars au 12 mai 2016 (Ontario) Du 8 mars au 4 mai 2016 (toutes les autres provinces)	434	Oui (26 avril 2016)
4.2, 5.2 et 6.2	Du 13 mai 2016 à aujourd'hui (Ontario) Du 5 mai 2016 à aujourd'hui (toutes les autres provinces)	399	Oui (6 juin 2016)
Total		**1 207**	

* Remarque : Une fois que les données relatives aux versions 4.1, 5.1 et 6.1 ont été fournies à Yardstick à des fins d'analyse, les examens ont continué à être utilisés jusqu'au début mai. Par conséquent, les données relatives à ces versions sont plus nombreuses que ce qu'indique le tableau ci-dessus.

Des analyses d'items et d'examen ont été réalisées pour deux versions des formats d'examens équivalents du PQAP (c.-à-d. les versions 4.1, 5.1 et 6.1 et les versions 4.2, 5.2 et 6.2). Les tableaux 2 à 5, ci-dessous, renferment un résumé des résultats associés aux versions 4.1, 5.1 et 6.1 de l'examen. Ces versions correspondent aux trois formats équivalents rattachés à chacun des examens modulaires suivants : *Déontologie – Common law, Assurance-vie, Assurance contre la maladie ou les accidents* et *Fonds distincts.*

Chaque tableau contient une estimation de la fiabilité de l'examen (c.-à-d. l'alpha de Cronbach et l'erreur standard de la mesure), des statistiques descriptives des scores (c.-à-d. le score moyen, minimum et maximum ainsi qu'un écart-type des scores), des statistiques descriptives des valeurs de p (c.-à-d. la valeur de p moyenne, minimum et maximum ainsi qu'un écart-type des valeurs de p) de même que des statistiques descriptives des coefficients de corrélation bisériale de point corrigés (c.-à-d. le coefficient de corrélation bisériale de point corrigé moyen, minimum et maximum ainsi qu'un écart-type du coefficient de corrélation bisériale de point corrigé).

D'après les résultats de l'*examen du module Déontologie – Common law*, les formats d'examens équivalents étaient semblables quant à la difficulté moyenne des items (valeur de p = 0,73, 0,73, et 0,70 pour les versions 4.1, 5.1 et 6.1 de l'examen, respectivement) et la capacité moyenne des items à établir une distinction entre les candidats forts et les candidats faibles (rpbis = 0,21, 0,21 et 0,17 pour les versions 4.1, 5.1 et 6.1 de l'examen, respectivement). Comme le montre le tableau 2, les versions 4.1 et 5.1 de l'*examen du module Déontologie – Common law* ne renfermaient pas d'items excessivement difficiles. En revanche, la version 6.1 en contenait un. En moyenne, les items des trois formats d'examens équivalents avaient un bon pouvoir de discrimination.

La fiabilité d'un examen dépend de sa longueur et de l'intervalle de la performance des candidats. Plus l'examen est court et plus l'intervalle des scores obtenus est étroit, moins l'examen est fiable. L'*examen du module Déontologie – Common law* est assez court : il compte 20 items. On ne peut donc pas s'attendre à ce que son degré de fiabilité soit élevé. Par conséquent, la faiblesse des coefficients de fiabilité des formats équivalents de l'*examen du module Déontologie – Common law* n'était pas surprenante.

Tableau 2. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Déontologie – Common law
(4.1, 5.1 et 6.1, versions anglaises seulement)

	Version 4.1	**Version 5.1**	**Version 6.1**
Nombre de candidats	455	410	423
QCM (n = 20)			
Moyenne	14,52	14,55	14,08
ET	2,86	2,84	2,60
Min.	6	5	3
Max.	20	20	20
Valeur de p des QCM			
Moyenne	0,73	0,73	0,70
ET	0,18	0,20	0,21
Min.	0,36	0,39	0,13
Max.	0,94	0,98	0,94
Rpbis des QCM			
Moyenne	0,21	0,21	0,17
ET	0,07	0,09	0,10
Min.	0,07	0,04	0,01
Max.	0,36	0,36	0,35
Alpha de Cronbach	0,62	0,63	0,53
ESM	1,76	1,73	1,78

Le tableau 3 renferme un résumé des résultats de l'analyse d'items et d'examen des versions 4.1, 5.1 et 6.1 des formats de l'*examen du module Assurance-vie*. Tous les formats d'examens se sont révélés semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,66, 0,66 et 0,67 et rpbis = 0,18, 0,21 et 0,27 pour les versions 4.1, 5.1 et 6.1 de l'examen, respectivement). La version 5.1 de l'examen ne contenait qu'un item difficile (valeur de p = 0,29); c'était aussi le cas de la version 6.1 (valeur de p = 0,30). Les indices moyens de discrimination des items des trois formats étaient acceptables, et la fiabilité de chaque format se situait entre 0,60 et 0,70.

Tableau 3. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Assurance-vie (4.1, 5.1 et 6.1, versions anglaises seulement)

	Version 4.1	**Version 5.1**	**Version 6.1**
Nombre de candidats	450	429	402
QCM (n = 30)			
Moyenne	19,89	19,64	20,18
ET	3,87	4,14	4,77
Min.	5	4	6
Max.	29	30	29
Valeur de p des QCM			
Moyenne	0,66	0,66	0,67
ET	0,16	0,18	0,16
Min.	0,35	0,29	0,30
Max.	0,91	0,97	0,98
Rpbis des QCM			
Moyenne	0,18	0,21	0,27
ET	0,09	0,10	0,09
Min.	0,00	-0,05	0,12
Max.	0,31	0,38	0,50
Alpha de Cronbach	0,62	0,68	0,77
ESM	2,39	2,34	2,29

Le tableau 4 renferme un résumé des résultats de l'analyse d'items et d'examen des versions 4.1, 5.1 et 6.1 des formats de l'*examen du module Assurance contre la maladie ou les accidents*. Tous les formats d'examens se sont révélés semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,70, 0,70 et 0,73 et rpbis = 0,22, 0,21 et 0,22 pour les versions 4.1, 5.1 et 6.1 de l'examen, respectivement). La version 5.1 de l'examen contenait deux items difficiles (valeur de p = 0,16 et 0,22), tandis que la version 6.1 n'en contenait qu'un (valeur de p = 0,24). Les indices moyens de discrimination des items des trois formats étaient acceptables, et la fiabilité de chaque format avoisinait 0,70.

Tableau 4. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Assurance contre la maladie ou les accidents (4.1, 5.1 et 6.1, versions anglaises seulement)

	Version 4.1	**Version 5.1**	**Version 6.1**
Nombre de candidats	466	445	433
QCM (n = 30)			
Moyenne	21,04	20,87	21,79

ET	4,12	3,76	3,90
Min.	6	9	9
Max.	30	29	30
Valeur de p des QCM			
Moyenne	0,70	0,70	0,73
ET	0,17	0,22	0,17
Min.	0,35	0,16	0,24
Max.	0,97	0,98	0,97
Rpbis des QCM			
Moyenne	0,22	0,21	0,22
ET	0,11	0,09	0,11
Min.	-0,13	0,05	-0,02
Max.	0,45	0,36	0,40
Alpha de Cronbach	0,70	0,67	0,69
ESM	2,26	2,16	2,17

Le tableau 5 renferme un résumé de l'analyse d'items et d'examen des versions 4.1, 5.1 et 6.1 des formats de l'*examen du module Fonds distincts*. Tous les formats d'examens se sont révélés semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,62, 0,62 et 0,62 et rpbis = 0,26, 0,24 et 0,27 pour les versions 4.1, 5.1 et 6.1 de l'examen, respectivement). La version 4.1 de l'examen ne contenait qu'un item difficile (valeur de p = 0,28). Les indices moyens de discrimination des items des trois formats étaient acceptables, et la fiabilité de chaque format dépassait 0,70.

Tableau 5. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Fonds distincts (4.1, 5.1 et 6.1, versions anglaises seulement)

	Version 4.1	**Version 5.1**	**Version 6.1**
Nombre de candidats	443	442	409
QCM (n = 30)			
Moyenne	18,47	18,56	18,53
ET	4,93	4,71	5,08
Min.	7	4	6
Max.	30	30	30
Valeur de p des QCM			
Moyenne	0,62	0,62	0,62
ET	0,15	0,14	0,15
Min.	0,28	0,38	0,34
Max.	0,91	0,91	0,83
Rpbis des QCM			
Moyenne	0,26	0,24	0,27
ET	0,09	0,09	0,09
Min.	0,04	0,04	0,07
Max.	0,41	0,41	0,41
Alpha de Cronbach	0,76	0,73	0,78
ESM	2,42	2,45	2,38

Les tableaux 6 à 9, ci-dessous, renferment un résumé des résultats statistiques associés aux versions 4.2, 5.2 et 6.2 de l'examen. Ce sont ces formats d'examens qui sont soumis aux candidats à l'heure actuelle.

Les résultats associés aux versions actuelles des formats de *l'examen du module Déontologie – Common law* sont très semblables à ceux qui se rapportaient aux versions précédentes. Les trois formats d'examens équivalents sont semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,71, 0,72 et 0,70 et rpbis = 0,21, 0,24 et 0,20 pour les versions 4.2, 5.2 et 6.2 de l'examen, respectivement). Aucun des formats ne contenait d'items très difficiles, ce qui constitue une amélioration par rapport aux précédentes versions de l'examen. En moyenne, les items avaient un bon pouvoir de discrimination. La fiabilité des versions actuelles de l'examen est semblable à celle des versions précédentes. Il convient de noter que la fiabilité du format 6.2 s'est améliorée.

Tableau 6. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Déontologie – Common law
(4.2, 5.2 et 6.2, versions anglaises seulement)

	Version 4.2	**Version 5.2**	**Version 6.2**
Nombre de candidats	422	375	395
QCM (n = 20)			
Moyenne	14,19	14,45	13,96
ET	2,88	2,98	2,98
Min.	6	3	3
Max.	20	20	20
Valeur de p des QCM			
Moyenne	0,71	0,72	0,70
ET	0,18	0,20	0,16
Min.	0,38	0,36	0,39
Max.	0,95	0,97	0,91
Rpbis des QCM			
Moyenne	0,21	0,24	0,20
ET	0,08	0,06	0,09
Min.	0,02	0,13	0,05
Max.	0,36	0,35	0,33
Alpha de Cronbach	0,61	0,67	0,60
ESM	1,80	1,71	1,88

Le tableau 7 renferme un résumé des résultats statistiques associés aux versions actuelles des formats de *l'examen du module Assurance-vie*. Les trois formats d'examens sont semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,67, 0,66 et 0,66 et rpbis = 0,22, 0,22 et 0,22 pour les versions 4.2, 5.2 et 6.2 de l'examen, respectivement). Ces versions de l'examen ne contiennent aucun item excessivement difficile, ce qui constitue une amélioration par rapport aux versions précédentes. Les indices moyens de discrimination des items des trois formats sont acceptables. La fiabilité de chaque format se situe en outre au-dessus de 0,70, ce qui semble attribuable au remplacement de certaines questions.

Tableau 7. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Assurance-vie
(4.2, 5.2 et 6.2, versions anglaises seulement)

	Version 4.2	Version 5.2	Version 6.2
Nombre de candidats	385	392	393
QCM (n = 30)			
Moyenne	20,11	19,77	19,56
ET	4,33	4,38	4,43
Min.	6	6	6
Max.	29	29	29
Valeur de p des QCM			
Moyenne	0,67	0,66	0,66
ET	0,15	0,16	0,16
Min.	0,32	0,34	0,32
Max.	0,91	0,96	0,98
Rpbis des QCM			
Moyenne	0,22	0,22	0,22
ET	0,09	0,11	0,09
Min.	0,03	0,00	0,05
Max.	0,38	0,40	0,42
Alpha de Cronbach	0,70	0,71	0,70
ESM	2,37	2,36	2,42

Le tableau 8 renferme un résumé des résultats statistiques associés aux versions actuelles des formats de *l'examen du module Assurance contre la maladie ou les accidents*. Tous les formats d'examens sont semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,71, 0,68 et 0,69 et rpbis = 0,26, 0,25 et 0,25 pour les versions 4.2, 5.2 et 6.2 de l'examen, respectivement). La version 5.1 de l'examen ne contient qu'un item difficile (valeur de p = 0,23); c'est aussi le cas de la version 6.1 (valeur de p = 0,26). Les indices moyens de discrimination des items des trois formats sont excellents, et la fiabilité de chaque format d'examen dépasse largement 0,70.

Tableau 8. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Assurance contre la maladie ou les accidents
(4.2, 5.2 et 6.2, versions anglaises seulement)

	Version 4.2	Version 5.2	Version 6.2
Nombre de candidats	424	392	401
QCM (n = 30)			
Moyenne	21,33	20,45	20,62
ET	4,47	4,44	4,49
Min.	6	7	3
Max.	30	29	29
Valeur de p des QCM			
Moyenne	0,71	0,68	0,69
ET	0,16	0,19	0,17
Min.	0,35	0,23	0,26
Max.	0,98	0,96	0,98
Rpbis des QCM			
Moyenne	0,26	0,25	0,25
ET	0,10	0,09	0,11
Min.	0,02	0,12	-0,10

Max.	0,45	0,38	0,45
Alpha de Cronbach	0,75	0,74	0,74
ESM	2,23	2,27	2,29

Le tableau 9 renferme un résumé des résultats statistiques associés aux versions actuelles des formats de *l'examen du module Fonds distincts*. Tous les formats d'examens se sont révélés semblables quant à la difficulté et au pouvoir de discrimination moyens des items (valeur de p = 0,63, 0,65 et 0,63 et rpbis = 0,26, 0,25 et 0,26 pour les versions 4.2, 5.2 et 6.2 de l'examen, respectivement). Ces versions de l'examen ne contiennent aucun item excessivement difficile, ce qui constitue une amélioration par rapport aux versions précédentes. Les indices moyens de discrimination des items des trois formats sont excellents, et la fiabilité de chaque format dépasse 0,70.

Tableau 9. Résultats de l'analyse d'items et d'examen portant sur l'examen du module Fonds distincts (4.2, 5.2 et 6.2, versions anglaises seulement)

	Version 4.2	**Version 5.2**	**Version 6.2**
Nombre de candidats	429	356	418
QCM (n = 30)			
Moyenne	18,82	19,43	18,66
ET	4,92	4,90	5,03
Min.	7	6	4
Max.	29	30	29
Valeur de p des QCM			
Moyenne	0,63	0,65	0,63
ET	0,13	0,12	0,12
Min.	0,38	0,41	0,38
Max.	0,92	0,90	0,83
Rpbis des QCM			
Moyenne	0,26	0,25	0,26
ET	0,10	0,09	0,10
Min.	0,00	0,08	0,04
Max.	0,49	0,41	0,44
Alpha de Cronbach	0,76	0,75	0,76
ESM	2,41	2,45	2,46

À la lumière des résultats des analyses d'items et d'examen portant sur deux versions différentes de formats équivalents de l'examen du PQAP, il est permis de conclure que ces formats d'examens sont fiables. Ils sont en outre constitués d'items de qualité dont le niveau de difficulté est raisonnable, en plus d'avoir une solide capacité à distinguer les candidats forts des candidats faibles. La taille des échantillons des deux séries d'analyses est suffisamment large pour pouvoir tirer des conclusions fiables quant à la qualité des items et des examens. Les résultats des analyses statistiques portant sur les versions actuelles des examens sont semblables à ceux qui se rapportaient aux versions précédentes.

Résumé et recommandations

- ***Critères statistiques d'évaluation de l'examen***. Les OCRA ont mis en place un processus transitoire pour surveiller la performance et évaluer la qualité d'un examen et de ses items. Le fait que les OCRA se basent sur des statistiques issues de larges échantillons pour mettre à jour leurs formats d'examens constitue une pratique digne de mention. Yardstick recommande de

procéder à des analyses d'items et d'examen de façon régulière et d'ajuster, si nécessaire, les scores de tous les candidats d'une cohorte. Cette stratégie aidera les OCRA à faire en sorte que les scores d'un examen restent significatifs d'un candidat à l'autre. Quand commencera la phase de maintenance de l'examen du PQAP, il sera important de documenter les résultats des analyses statistiques réalisées sur tous les formats d'examens à chaque passation de l'examen.

- ***Résultats d'analyses d'items et d'examen.*** Yardstick a procédé à des analyses statistiques pour évaluer les propriétés psychométriques des versions précédentes et actuelles des formats d'examens du PQAP. Les résultats se sont révélés semblables entre les différentes versions, les plus récentes ayant des propriétés psychométriques légèrement plus favorables que les précédentes. Les items excessivement difficiles ont été éliminés de l'examen et remplacés par d'autres, moins ardus.

  La fiabilité des formats d'examens actuels se situe dans une plage acceptable compte tenu de leur petite taille. Les items ont un niveau de difficulté modéré et une solide capacité à distinguer les candidats forts des candidats faibles. Les formats des quatre examens ont une fiabilité semblable et des statistiques moyennes similaires pour leurs différents items, éléments qui attestent l'équivalence des formats en question. En conclusion, les formats d'examens équivalents sont comparables non seulement par leur contenu, mais aussi par leurs propriétés psychométriques.

Étape VI – Établissement de la norme

Dans le cadre de l'étape VI, Yardstick a passé en revue la documentation décrivant le processus utilisé pour établir une norme relative à l'examen.

Pour les organismes de certification, l'établissement d'une norme renvoie au fait d'appliquer une norme de performance (c.-à-d. un seuil de réussite) à un examen. Dans le cas des examens certificatifs, la norme de performance permet de diviser l'intervalle des scores de manière à ce que ceux-ci soient répartis dans deux catégories de performance à l'examen : la réussite ou l'échec. Les candidats qui réussissent l'examen (c.-à-d. qui obtiennent un score égal ou supérieur à la norme) sont jugés compétents : ils ont prouvé qu'ils ont les connaissances ou les compétences requises pour pratiquer leur profession de façon sûre et efficace. Les candidats qui échouent à l'examen (c.-à-d. qui obtiennent un score inférieur à la norme) sont jugés insuffisamment compétents : ils n'ont pas prouvé qu'ils ont les connaissances ou les compétences requises.

L'établissement d'une norme est une étape critique dans l'élaboration et l'utilisation d'examens certificatifs, car il permet aux associations professionnelles et aux organismes de réglementation de prendre des décisions appropriées sur les candidats. Ce processus vise à établir une norme de performance qui constitue une importante preuve de validité de l'interprétation des scores d'un examen.

Plusieurs normes de test régissent les processus d'établissement de normes dans les organismes de certification.
La *norme 17 de la NCCA* précise que la méthode employée pour fixer des normes de performance doit être basée sur des principes de mesure généralement acceptables qui cadrent avec l'objectif de l'examen. Par ailleurs, le processus d'établissement de norme devrait être documenté de façon assez détaillée pour pouvoir être reproduit : il devrait notamment comprendre une description de la méthode employée et de l'issue du processus. Selon la *norme 5.21*, quand une norme sert à classer les candidats dans des catégories de performance distinctes – la réussite ou l'échec, par exemple –, les procédures d'établissement de normes doivent être justifiées. Elles doivent en outre être documentées clairement et avec suffisamment de détails.

> ***Norme 5.21*** Lorsque les interprétations de scores suggérées sont associées à un ou à plusieurs seuils de réussite, la justification et les procédés utilisés pour les établir devraient être clairement documentés.

D'après la documentation fournie par les OCRA, il semble que le seuil de réussite de l'examen du PQAP ait été établi au moyen d'une approche normative. Les procédures d'établissement de normes se classent en deux grandes catégories : normatives et absolues (ou critériées). Les approches normatives reposent sur un seuil de réussite arbitraire fondé sur la performance d'un échantillon de candidats. Elles sont utilisées dans un contexte de recrutement de personnel, quand le nombre de postulants dépasse le nombre de postes à pourvoir. Les procédures d'établissement de normes absolues reposent quant à elles sur un seuil de réussite fondé sur les critères régissant une pratique professionnelle acceptable.

Le rapport technique préparé par un cabinet-conseil indépendant en 2001 décrit l'approche normative d'établissement de normes associée à l'examen du PQAP. L'établissement du seuil de réussite à l'aide de cette approche, au détriment de la procédure d'établissement de normes absolues, n'est cependant pas justifié dans le rapport. Les procédures d'établissement de normes absolues sont pourtant celles que

privilégient – et de loin – les organismes de certification, car elles tiennent compte des propriétés d'un examen, des caractéristiques de la population de candidats et, surtout, du niveau de performance requis pour qu'un candidat exerce ses activités professionnelles avec compétence.

Le rapport technique sur l'établissement d'une norme contenait trois options de calcul du seuil de réussite à l'aide des données d'un prétest. Certains éléments n'étaient toutefois pas précisés clairement, soit la manière dont le prétest a été effectué, qui étaient les participants au prétest, quels formats d'examens ont été soumis aux participants, quel groupe a été choisi comme groupe de référence ainsi que la performance obtenue à l'examen par ce groupe. En outre, le seuil de réussite recommandé n'était pas documenté dans le rapport final.

Dans un courriel, les OCRA ont fait savoir que le seuil de réussite de l'examen du PQAP a été fixé sur la base des données d'un prétest soumis à un groupe de professionnels de l'assurance de personnes ayant deux ans d'expérience dans le domaine. Le seuil de réussite est issu du score le plus faible obtenu à l'examen par un membre de ce groupe : la limite inférieure de la plage de résultats a été calculée par soustraction à partir de ce score. En janvier 2016, les OCRA ont commencé à appliquer le même seuil de réussite « historique » à tous les examens modulaires. Pour réussir l'examen, un candidat doit obtenir au moins la note de passage à tous les examens modulaires. De même, il semble que pendant la période de transition, les OCRA abaissent le seuil de réussite d'un ou deux points les semaines où le taux de réussite est inférieur à 70 %.

Compte tenu du fait que l'examen du PQAP est un examen d'accréditation qui vise à reconnaître la compétence chez les professionnels de l'assurance de personnes, le seuil de réussite de l'examen devrait refléter une norme de compétence établie par des experts. Il serait donc préférable de miser sur une procédure d'établissement de normes absolues : avec une telle méthode, le seuil de réussite repose sur le niveau de performance nécessaire pour qu'un candidat exerce sa profession de façon sûre et compétente.

Le fait d'appliquer une approche normative d'établissement de normes aux examens certificatifs soulève plusieurs préoccupations. Tout d'abord, les standards normatifs garantissent que certains candidats échoueront à l'examen, peu importe les connaissances qu'ils ont démontrées. Si le standard se situe à un écart-type sous la moyenne, il est assuré que 16 % de la population de candidats échoueront à l'examen, quelles que soient leurs connaissances. Dans le cas de l'examen du PQAP, il n'est pas établi clairement pourquoi le point de référence de l'établissement de la norme réside dans la performance du candidat le plus faible au sein d'un groupe de professionnels de l'assurance de personnes ayant deux ans d'expérience. Ce choix est d'autant plus préoccupant que les participants au prétest ne représentent pas nécessairement la population cible de candidats passant cet examen, en raison d'un manque de formation pertinente.

Yardstick estime que la méthode Angoff modifiée est la méthode d'établissement de normes qui convient le mieux à un examen d'accréditation. Diverses méthodes permettent d'établir des normes de performance à un examen. La méthode Angoff (et ses dérivés) est l'une des plus couramment utilisées pour établir des normes de performance dans le cadre d'examens certificatifs à choix multiple. Dans un sondage sur l'établissement de normes mené en 2008 auprès d'organismes offrant des programmes de certification agréés par la NCCA, 75 % des organismes utilisaient la méthode Angoff modifiée. Quand ils ont été invités à citer les facteurs ayant le plus d'influence sur leur choix de méthode d'établissement de normes, les organismes ont mentionné la fiabilité de la méthode, sa conformité aux normes d'agrément de la NCCA ou de l'ISO/CEI et sa facilité d'utilisation.

L'autre préoccupation soulevée par l'emploi de l'approche normative d'établissement de normes réside dans le manque de sensibilité de cette méthode aux changements dans le niveau de compétence des

candidats au fil du temps. Cette caractéristique peut mener à la création de groupes de professionnels certifiés n'ayant pas tous le même niveau de compétence. De plus, si les critères de compétence qui prévalent dans l'industrie sont devenus plus stricts au cours de la dernière décennie, le seuil de réussite actuel de l'examen du PQAP ne reflétera pas ce changement parce que ce seuil n'a jamais été lié à quelque norme de performance que ce soit. Comme le seuil de réussite actuel n'est pas explicitement associé à la norme de compétence attendue d'un professionnel en début de carrière, il ne peut soutenir la validité des conclusions tirées à partir du score de l'examen.

Une fois qu'un seuil de réussite a été fixé, il est considéré comme une valeur fixe jusqu'à ce qu'un groupe de décideurs estime qu'il ne convient plus. La décision de modifier le seuil de réussite n'est jamais prise à la légère. Elle est prise par le groupe de décideurs au moyen d'une erreur standard de la moyenne, laquelle est calculée à partir des cotes établies dans le cadre de l'établissement d'une norme. Par exemple, le groupe susmentionné pourrait décider de hausser ou d'abaisser le seuil de réussite à un examen en additionnant ou en soustrayant une ou deux erreurs standard de la moyenne de ces cotes. L'erreur standard de la moyenne correspond au degré de changement dans le seuil de réussite auquel on peut s'attendre en répliquant l'étude d'établissement de normes. Il convient de noter qu'il est impossible de procéder à un ajustement valide du seuil de réussite sans réaliser tout d'abord une étude d'établissement de norme.

Les décideurs doivent avoir une raison sérieuse de modifier le seuil de réussite. Par exemple, ils peuvent décider que lorsqu'un candidat se trouve à la limite du seuil de réussite à l'examen, il est préférable de lui attribuer une réussite plutôt qu'un échec. Avant de prendre cette décision, les décideurs compareront les conséquences d'une diminution du seuil de réussite avec celles liées au fait de le laisser tel quel. Par exemple, abaisser le seuil de réussite peut augmenter le nombre de faux positifs, c'est-à-dire les candidats qui réussissent l'examen, mais qui, dans les faits, sont incompétents et doivent leur réussite à la chance. Si ces candidats réussissent l'examen, ils obtiendront leur certification. Les décideurs doivent tenir compte de ce risque pour déterminer s'il est justifié d'abaisser le seuil de réussite.

Selon la *norme 11.16*, le seuil de réussite à un examen doit dépendre de la norme de performance (c.-à-d. des connaissances et des compétences requises pour qu'un candidat exerce ses activités professionnelles avec compétence) plutôt que du taux de réussite de l'examen. Le seuil de réussite ne doit pas être ajusté pour réguler le nombre de candidats qui réussissent l'examen du PQAP.

> ***Norme 11.16*** Le niveau de performance exigé pour réussir un test certificatif devrait dépendre des connaissances et des compétences nécessaires pour fournir une performance acceptable dans le cadre du poste ou de l'exercice de la profession. De même, il ne devrait pas être ajusté dans le seul but de réguler le nombre ou la proportion de personnes qui réussissent le test.

Selon la *norme 17 de la NCCA*, le programme de certification devrait évaluer les normes de compétence suffisamment souvent pour qu'elles suivent l'évolution de la pratique professionnelle. Tout changement significatif au champ d'exercice ou aux spécifications d'un examen justifie la tenue d'une nouvelle étude d'établissement de norme. Compte tenu du fait qu'un nouveau *Profil de compétences – Représentant en assurance de personnes* et de nouvelles spécifications d'examen ont été élaborés récemment, il ne fait nul doute que l'examen du PQAP doit faire l'objet d'une nouvelle étude d'établissement de norme.

Résumé et recommandations

- ***Méthode d'établissement d'une norme.*** Yardstick recommande que les OCRA mènent une nouvelle étude d'établissement de norme pour que le seuil de réussite de l'examen du PQAP soit lié à la norme de performance absolue à laquelle doivent se conformer les représentants en assurance de personnes qui amorcent une carrière dans l'industrie. L'utilisation d'un seuil de réussite « historique » par les OCRA est problématique pour plusieurs raisons, et la récente modification du champ d'exercice des représentants en assurance de personnes justifie la tenue d'une nouvelle étude d'établissement de norme. Yardstick recommande que les OCRA utilisent la méthode Angoff modifiée pour établir un seuil de réussite à chacun des modules de l'examen du PQAP. En vertu de cette méthode, un groupe d'experts de contenu doit conceptualiser le niveau de performance auquel on peut s'attendre d'un candidat minimalement compétent et appliquer ce point de référence à l'évaluation des questions de l'examen. Pour chaque question, les normalisateurs estiment le pourcentage de candidats minimalement compétents qui répondront correctement. Avant de rendre leurs jugements, les normalisateurs reçoivent une formation structurée et discutent des attentes qu'ils entretiennent envers les candidats minimalement compétents. Les normalisateurs sont appelés à rendre leurs jugements en deux rondes, lesquelles sont entrecoupées d'une discussion avec des pairs et de la présentation de données empiriques sur le niveau de difficulté des items. Dans le cadre de la séance de discussion et de la présentation de données empiriques, les normalisateurs obtiennent une rétroaction qui est censée accroître la justesse de leurs jugements. L'application de la méthode Angoff modifiée aidera les OCRA à établir un seuil de réussite distinct pour chaque format d'examen, ce qui contribuera à établir l'équivalence de ces formats et à soutenir la validité de l'interprétation des scores de l'examen. Ces seuils de réussite seront étroitement liés aux normes de compétence de la profession et refléteront la difficulté des questions d'un format d'examen en particulier.

- ***Documentation de la méthode d'établissement d'une norme et des résultats***. Le rapport technique sur l'établissement de normes publié en 2001 ne contenait pas assez de détails sur le processus employé et les résultats obtenus. Après avoir mené la nouvelle étude d'établissement d'une norme, il serait judicieux que les OCRA documentent les renseignements suivants dans le rapport relatif à l'étude : 1) la justification du choix de la méthode d'établissement d'une norme; 2) les procédures relatives à la sélection et à la formation des normalisateurs; 3) les qualifications des normalisateurs; 4) les méthodes d'établissement d'une norme; 5) une description conceptuelle du niveau de performance sous-jacent à la norme; 6) les activités de collecte de données; 7) les normes recommandées; et 8) les ajustements apportés par les décideurs à la norme établie, s'il y a lieu.

- ***Ajustement du seuil de réussite.*** Quand un seuil de réussite a été fixé, il ne peut être ajusté que par un groupe de décideurs ayant des raisons sérieuses de croire qu'il a été établi de façon inappropriée ou qu'il n'est plus adéquat. Un faible taux de réussite peut justifier l'ajustement de ce seuil. Cela dit, une fois qu'un nouveau seuil de réussite est établi, il ne peut être modifié librement pour réguler le taux de réussite à l'examen au fil du temps. L'ajustement de ce seuil n'est justifiable que s'il est basé sur des données statistiques obtenues dans le cadre de l'établissement d'une norme.

- ***Réévaluation du seuil de réussite.*** Une fois qu'un seuil de réussite a été défini, il doit être évalué de façon régulière pour déterminer s'il reflète toujours le niveau de performance nécessaire pour qu'un candidat exerce sa profession de façon sûre et compétente. Tout

changement au champ d’exercice des représentants en assurance de personnes ou à l’ensemble des connaissances que ces représentants doivent posséder devrait donner lieu à la révision et à l’ajustement du seuil de réussite, ce qui peut nécessiter la tenue d’une nouvelle étude d’établissement de norme.

Annexe A

Introduction

Cet addenda vise à commenter et à décrire les résultats des analyses de données additionnelles réalisées par Yardstick à la lumière des observations formulées par les intervenants des OCRA sur l'analyse psychométrique de l'examen du PQAP. La qualité de l'examen du PQAP est un élément incontournable pour évaluer, sur la base d'information significative, si les candidats sont prêts à exercer la profession de représentant en assurance de personnes. Il est donc important que toutes les parties concernées prennent connaissance des conclusions de l'analyse psychométrique et en discutent.

Les questions d'ordre psychométrique suivantes sont abordées dans cet addenda :

1) Incidence des examens modulaires du PQAP de 2016 sur le **niveau de difficulté et le taux de réussite de l'examen**;
2) Incidence de la **fiabilité de l'examen** sur les décisions prises relativement au classement des candidats;
3) Nécessité de tenir compte de l'incidence des variables démographiques, sociodémographiques et culturelles sur les résultats de l'examen – **équité**.

Taux de réussite et niveau de difficulté

Pour évaluer la possibilité que l'introduction des examens modulaires ait contribué au déclin du taux de réussite à l'examen, les OCRA ont demandé à Yardstick de comparer les taux de réussite et les niveaux de difficulté associés à l'ancien et au nouvel examen du PQAP. La comparaison des niveaux de difficulté de différentes versions de l'examen dépasse le cadre d'une analyse psychométrique normale, l'ancienne version n'étant plus utilisée.

Les OCRA ont fourni à Yardstick les taux de réussite nationaux aux examens modulaires depuis le début de l'année 2016 ainsi que les taux de réussite historiques à l'examen du PQAP entre 2007 et 2015. Ils ont aussi communiqué à Yardstick la valeur de p des items qui figuraient dans les trois formats de l'examen du PQAP au cours des six premiers mois de l'année 2015.

Taux de réussite

Yardstick a comparé les taux de réussite nationaux à l'ancien examen du PQAP avec les taux de réussite nationaux aux nouveaux examens modulaires. Les résultats de la comparaison des taux de réussite sont présentés aux tableaux 1 à 3.

Comme le montre le tableau 1, les taux de réussite historiques moyens à l'ancien examen du PQAP se situent entre 64,0 % et 74,6 %. En 2015, le taux de réussite à l'examen du PQAP était de 69,2 %, ce qui se situe dans la plage des taux de réussite historiques.

Tableau A1. Taux de réussite historiques à l'examen du PQAP (2007-2015)[1]

Année	N^bre d'examens administrés	Taux de réussite
2015	4 817	69,2 %
2014	11 067	72,5 %
2013	10 180	72,0 %
2012	9 501	71,9 %
2011	8 953	74,6 %
2010	9 250	67,2 %
2009	10 382	64,0 %
2008	9 993	64,7 %
2007	9 982	74,1 %

Les tableaux 2 et 3 font état des taux de réussite moyens aux trois formats des examens modulaires en 2016. Pour les formats 4.2, 5.2 et 6.2, les taux de réussite à l'examen modulaire varient entre 66,2 % et 84,0 % (tableau 2), tandis que le taux de réussite le plus élevé et le taux de réussite le plus faible à l'examen modulaire sont respectivement de 69,6 % et de 84,6 % (tableau 3). En outre, une légère tendance à la hausse peut être observée dans les taux de réussite nationaux après l'introduction des formats 4.3, 5.3 et 6.3 de l'examen.

Tableau A2. Taux de réussite à l'examen du PQAP en 2016 (janvier-juin 2016)

Module de l'examen	N^o du format	N^bre d'examens administrés	Taux de réussite
Déontologie – Common law	4.2	987	82,2 %
	5.2	910	84,7 %
	6.2	947	85,2 %
	Total	**2 844**	**84,0 %**
Assurance-vie	4.2	1 137	77,5 %
	5.2	1 102	71,9 %
	6.2	1 107	73,1 %
	Total	**3 346**	**74,2 %**

[1] Au tableau 1, les taux de réussite à l'examen de 2007 à 2014 sont tirés du site Web des OCRA. Le taux de réussite à l'examen de 2015 est calculé d'après les trois formats d'examens utilisés dans l'ensemble des provinces, sauf en Colombie-Britannique et au Québec, au cours des six premiers mois de l'année 2015. Ce sont ces trois formats qui ont été utilisés aux fins de la comparaison de l'ancienne et de la nouvelle version de l'examen dans cette publication.

Assurance contre la maladie ou les accidents	4.2	1 195	82,0 %
	5.2	1 122	79,2 %
	6.2	1 149	81,1 %
	Total	**3 466**	**80,8 %**
Fonds distincts	4.2	1 205	64,0 %
	5.2	1 089	67,7 %
	6.2	1 183	66,9 %
	Total	**3 477**	**66,2 %**

Tableau A3. Taux de réussite à l'examen du PQAP en 2016 (juillet 2016)

Module de l'examen	N° du format	Nbre d'examens administrés	Taux de réussite
Déontologie – Common law	4.3	456	83,1 %
	5.3	425	86,1 %
	6.3	432	84,5 %
	Total	**1 313**	**84,6 %**
Assurance-vie	4.3	523	78,2 %
	5.3	506	78,1 %
	6.3	492	73,8 %
	Total	**1 521**	**76,7 %**
Assurance contre la maladie ou les accidents	4.3	545	80,4 %
	5.3	516	80,6 %
	6.3	521	83,1 %
	Total	**1 582**	**81,4 %**
Fonds distincts	4.3	560	67,7 %
	5.3	525	69,3 %
	6.3	509	71,7 %
	Total	**1 594**	**69,6 %**

Comme l'indiquent les tableaux récapitulatifs ci-dessus, les taux de réussite au nouvel examen modulaire sont légèrement plus élevés que les taux de réussite à l'ancien examen, sauf pour le module *Fonds distincts*. Néanmoins, en juillet 2016, 69,6 % des candidats ont réussi le nouvel examen du module *Fonds distincts* : il s'agit d'une légère amélioration par rapport au taux de réussite à l'examen du PQAP en 2015, qui se situait alors à 69,2 %.

Mentionnons également que le taux de réussite moyen pour l'ensemble des modules de l'examen du PQAP en 2016 est de 78,1 %. Ce taux de réussite est supérieur à tous les taux de réussite obtenus à l'examen du PQAP dans le passé.

En somme, **les candidats réussissent mieux aux examens modulaires de 2016 du PQAP qu'aux anciens examens du PQAP.** La comparaison des taux de réussite nationaux aux anciens examens et

aux nouveaux examens modulaires a révélé que la notation conjonctive (c.-à-d. la notation par module) n'a pas eu de répercussions négatives sur les taux de réussite à l'examen.

Niveau de difficulté

Yardstick a comparé le niveau de difficulté des examens de 2015 et de 2016 du PQAP. Elle a plus particulièrement comparé la valeur de p associée aux items de l'examen de 2015 du PQAP avec celle associée aux items des deux plus récentes versions des examens modulaires de 2016 du PQAP. L'examen de 2015 du PQAP comportait 140 items, tandis que les examens modulaires de 2016 du PQAP en contenaient 20 ou 30 chacun. Il convient de mentionner qu'il existe des différences significatives entre la structure de l'ancien examen et celle des nouveaux examens modulaires. Le contenu de chaque nouvel examen modulaire ne correspond qu'à une partie de l'ancien examen, dont la portée était plus vaste. L'ancien examen contenait un certain nombre d'items qui mesuraient simultanément des compétences associées à plusieurs examens modulaires. Il n'a donc pas été possible de diviser nettement l'ancien examen en sections correspondant aux examens modulaires actuels; c'est d'ailleurs pourquoi ce dernier a été considéré comme un tout dans les analyses subséquentes. Yardstick reconnaît qu'en divisant l'ancien examen en sections correspondant aux compétences, il aurait été possible de comparer de façon plus directe le niveau de difficulté des items de l'ancien et du nouvel examen.

Le tableau 4 renferme un résumé de la valeur de p des items des trois formats de l'examen de 2015 du PQAP. En moyenne, de 62 à 64 % des candidats ont répondu correctement aux items de l'examen du PQAP de 2015. La vaste majorité des items (de 90 à 95 %) ont donné lieu à une réponse correcte par plus du tiers des candidats. Seulement 1 ou 2 % des items de l'examen, qui en compte 140, ont été jugés très difficiles (c.-à-d. que moins du tiers des candidats y ont répondu correctement). Ces conclusions tendent à indiquer que l'examen du PQAP de 2015 avait un degré de difficulté modéré pour les candidats.

Tableau A4. Valeur de p des items de l'examen de 2015 du PQAP

	Format 530	**Format 531**	**Format 532**
Valeur de p minimum	0,23	0,23	0,23
Valeur de p maximum	0,92	0,97	0,97
Valeur de p moyenne	0,62	0,64	0,64
ET des valeurs de p	0,16	0,17	0,17
N^{bre} de valeurs de p entre 0,20 et 0,29	2 (1,4 %)	3 (2,0 %)	2 (1,4 %)
N^{bre} de valeurs de p entre 0,30 et 0,59	58 (41,4 %)	56 (40,0 %)	59 (42,1 %)
N^{bre} de valeurs de p entre 0,60 et 0,89	76 (54,3 %)	70 (50,0 %)	69 (49,3 %)
N^{bre} de valeurs de p entre 0,90 et 1,00	4 (2,9 %)	11 (8,0 %)	10 (7,1 %)
N^{bre} total	140 (100 %)	140 (100 %)	140 (100 %)

Remarque : N (format 530) = 1 689, N (format 531) = 1 623 et N (format 532) = 1 505.

Les tableaux 5 à 8 renferment un résumé de la valeur de p des items des formats 4.2, 5.2 et 6.2 de l'examen du PQAP de 2016. Ces formats ont été examinés dans le rapport d'analyse psychométrique. En moyenne, de 70 à 72 % des candidats ont répondu correctement aux items de l'examen de 2016 du module *Déontologie – Common law*; de 66 à 67 % ont répondu correctement aux items de l'examen du

module *Assurance-vie*; de 68 à 71 % ont répondu correctement aux items de l'examen du module *Assurance contre la maladie ou les accidents*; et finalement, de 63 à 65 % ont répondu correctement aux items de l'examen du module *Fonds distincts*. D'après le niveau moyen de difficulté des items, le nouvel examen était un peu plus facile que l'ancien.

La répartition des items des examens modulaires du PQAP de 2016 en fonction de leur valeur de p était semblable à celle de l'examen du PQAP de 2015. La plupart des items ont donné lieu à une réponse correcte par 30 à 90 % des candidats. Le pourcentage d'items auxquels 60 à 89 % des candidats ont répondu correctement a augmenté par rapport à l'ancien examen. De plus, la répartition des valeurs de p présentait une asymétrie négative plus marquée dans le nouvel examen que dans l'ancien. **Toutes ces données tendent à indiquer que les examens modulaires du PQAP de 2016 (formats 4.2, 5.2 et 6.2) étaient légèrement plus faciles pour les candidats que l'ancien examen.**

Tableau A5. Valeur de p des items de l'examen de 2016 du module Déontologie – Common law (formats 4.2, 5.2 et 6.2)

	Format 4.2	**Format 5.2**	**Format 6.2**
Valeur de p minimum	0,38	0,36	0,39
Valeur de p maximum	0,95	0,97	0,91
Valeur de p moyenne	0,71	0,72	0,70
ET des valeurs de p	0,18	0,20	0,16
N^{bre} de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N^{bre} de valeurs de p entre 0,30 et 0,59	7 (35,0 %)	5 (25,0 %)	5 (25,0 %)
N^{bre} de valeurs de p entre 0,60 et 0,89	11 (50,0 %)	12 (60,0 %)	14 (70,0 %)
N^{bre} de valeurs de p entre 0,90 et 1,00	2 (10,0 %)	3 (15,0 %)	1 (5,0 %)
Total	20 (100 %)	20 (100 %)	20 (100 %)

Remarque : N (format 4.2) = 422, N (format 5.2) = 375 et N (format 6.2) = 395.

Tableau A6. Valeur de p des items de l'examen de 2016 du module Assurance-vie (formats 4.2, 5.2 et 6.2)

	Format 4.2	**Format 5.2**	**Format 6.2**
Valeur de p minimum	0,32	0,34	0,32
Valeur de p maximum	0,91	0,96	0,98
Valeur de p moyenne	0,67	0,66	0,66
ET des valeurs de p	0,15	0,16	0,16
N^{bre} de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N^{bre} de valeurs de p entre 0,30 et 0,59	11 (36,7 %)	11 (36,6 %)	11 (36,6 %)
N^{bre} de valeurs de p entre 0,60 et 0,89	18 (60,0 %)	16 (53,3 %)	17 (56,7 %)
N^{bre} de valeurs de p entre 0,90 et 1,00	1 (3,3 %)	3 (10 %)	2 (6,7 %)
Total	30 (100 %)	30 (100 %)	30 (100 %)

Remarque : N (format 4.2) = 385, N (format 5.2) = 392 et N (format 6.2) = 393.

Tableau A7. Valeur de p des items de l'examen de 2016 du module Assurance contre la maladie ou les accidents (formats 4.2, 5.2 et 6.2)

	Format 4.2	Format 5.2	Format 6.2
Valeur de p minimum	0,35	0,23	0,26
Valeur de p maximum	0,98	0,96	0,98
Valeur de p moyenne	0,71	0,68	0,69
ET des valeurs de p	0,16	0,19	0,17
N[bre] de valeurs de p entre 0,20 et 0,29	0 (0 %)	1 (3,3 %)	1 (3,3 %)
N[bre] de valeurs de p entre 0,30 et 0,59	7 (23,3 %)	6 (20,0 %)	9 (30 %)
N[bre] de valeurs de p entre 0,60 et 0,89	21 (70,0 %)	19 (63,3 %)	17 (56,7 %)
N[bre] de valeurs de p entre 0,90 et 1,00	2 (6,7 %)	4 (13,3 %)	3 (10,0 %)
Total	30 (100 %)	30 (100 %)	30 (100 %)

Remarque : N (format 4.2) = 424, N (format 5.2) = 392 et N (format 6.2) = 401.

Tableau A8. Valeur de p des items de l'examen de 2016 du module Fonds distincts (formats 4.2, 5.2 et 6.2)

	Format 4.2	Format 5.2	Format 6.2
Valeur de p minimum	0,38	0,41	0,38
Valeur de p maximum	0,92	0,90	0,83
Valeur de p moyenne	0,63	0,65	0,63
ET des valeurs de p	0,13	0,12	0,12
N[bre] de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N[bre] de valeurs de p entre 0,30 et 0,59	12 (40,0 %)	10 (33,3 %)	10 (33,3 %)
N[bre] de valeurs de p entre 0,60 et 0,89	17 (56,7 %)	19 (63,3 %)	20 (66,7 %)
N[bre] de valeurs de p entre 0,90 et 1,00	1 (3,3 %)	1 (3,3 %)	0 (0 %)
Total	30 (100 %)	30 (100 %)	30 (100 %)

Remarque : N (format 4.2) = 429, N (format 5.2) = 356 et N (format 6.2) = 418.

En juillet 2016, les OCRA ont remplacé les formats d'examens 4.2, 5.2 et 6.2 par les formats 4.3, 5.3 et 6.3. Le tableau 9 renferme un résumé de la valeur de p des items des nouveaux formats d'examens. En moyenne, de 71 à 73 % des candidats ont répondu correctement aux items de l'examen du module *Déontologie – Common law*; de 68 à 69 % ont répondu correctement aux items de l'examen du module *Assurance-vie*; de 71 à 73 % ont répondu correctement aux items de l'examen du module *Assurance contre la maladie ou les accidents*; et de 65 à 68 % ont répondu correctement aux items de l'examen du module *Fonds distincts*. Le niveau de difficulté moyen des formats d'examens 4.3, 5.3 et 6.3 était très semblable à celui des formats d'examens 4.2, 5.2 et 6.2.

La répartition des items des examens modulaires en fonction de leur valeur de p était semblable dans les deux versions de l'examen du PQAP de 2016. Il convient en outre de noter que les OCRA se sont efforcés d'éliminer les items excessivement difficiles des formats d'examens. L'ancien examen du PQAP comportait deux ou trois items auxquels moins du tiers des candidats pouvaient répondre

correctement, tandis que la plupart des nouveaux examens du PQAP ne comportent pas d'items du genre.

Tableau A9. Valeur de p des items de l'examen de 2016 du module Déontologie – Common law (formats 4.3, 5.3 et 6.3)

	Format 4.3	Format 5.3	Format 6.3
Valeur de p minimum	0,42	0,38	0,38
Valeur de p maximum	0,95	0,97	0,93
Valeur de p moyenne	0,72	0,73	0,71
ET des valeurs de p	0,18	0,19	0,16
N[bre] de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N[bre] de valeurs de p entre 0,30 et 0,59	8 (40,0 %)	5 (25,0 %)	5 (25,0 %)
N[bre] de valeurs de p entre 0,60 et 0,89	10 (50,0 %)	11 (55,0 %)	14 (70,0 %)
N[bre] de valeurs de p entre 0,90 et 1,00	2 (10,0 %)	4 (20,0 %)	1 (5,0 %)
Total	20 (100 %)	20 (100 %)	20 (100 %)

Remarque : N (4.3) = 452, N (5.3) = 421 et N (6.3) = 429.

Tableau A10. Valeur de p des items de l'examen de 2016 du module Assurance-vie (formats 4.3, 5.3 et 6.3)

	Format 4.3	Format 5.3	Format 6.3
Valeur de p minimum	0,49	0,42	0,45
Valeur de p maximum	0,97	0,96	0,97
Valeur de p moyenne	0,69	0,69	0,68
ET des valeurs de p	0,15	0,15	0,13
N[bre] de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N[bre] de valeurs de p entre 0,30 et 0,59	12 (40,0 %)	10 (33,3 %)	10 (33,3 %)
N[bre] de valeurs de p entre 0,60 et 0,89	16 (53,3 %)	17 (56,7 %)	18 (60,0 %)
N[bre] de valeurs de p entre 0,90 et 1,00	2 (6,7 %)	3 (10,0 %)	2 (6,7 %)
Total	30 (100 %)	30 (100 %)	30 (100 %)

Remarque : N (4.3) = 451, N (5.3) = 439 et N (6.3) = 428.

Tableau A11. Valeur de p des items de l'examen de 2016 du module Assurance contre la maladie ou les accidents (formats 4.3, 5.3 et 6.3)

	Format 4.3	Format 5.3	Format 6.3
Valeur de p minimum	0,32	0,35	0,42
Valeur de p maximum	0,98	0,97	0,98
Valeur de p moyenne	0,71	0,72	0,73
ET des valeurs de p	0,17	0,17	0,14
N[bre] de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N[bre] de valeurs de p entre 0,30 et 0,59	5 (16,7 %)	7 (23,3 %)	4 (13,3 %)

N[bre] de valeurs de p entre 0,60 et 0,89	23 (76,7 %)	20 (66,7 %)	22 (73,3 %)
N[bre] de valeurs de p entre 0,90 et 1,00	2 (6,7 %)	3 (10,0 %)	4 (13,3 %)
Total	30 (100 %)	30 (100 %)	30 (100 %)

Remarque : N (4.3) = 476, N (5.3) = 441 et N (6.3) = 450.

Tableau A12. Valeur de p des items de l'examen de 2016 du module Fonds distincts (formats 4.3, 5.3 et 6.3)

	Format 4.3	**Format 5.3**	**Format 6.3**
Valeur de p minimum	0,43	0,43	0,50
Valeur de p maximum	0,90	0,91	0,87
Valeur de p moyenne	0,66	0,65	0,68
ET des valeurs de p	0,13	0,12	0,11
N[bre] de valeurs de p entre 0,20 et 0,29	0 (0 %)	0 (0 %)	0 (0 %)
N[bre] de valeurs de p entre 0,30 et 0,59	11 (36,7 %)	8 (26,7 %)	9 (30,0 %)
N[bre] de valeurs de p entre 0,60 et 0,89	19 (63,3 %)	21 (70,0 %)	21 (70,0 %)
N[bre] de valeurs de p entre 0,90 et 1,00	0 (0 %)	1 (3,3 %)	0 (0 %)
Total	30 (100 %)	30 (100 %)	30 (100 %)

Remarque : N (4.3) = 487, N (5.3) = 451 et N (6.3) = 442.

D'après la comparaison des valeurs moyennes et la répartition des indices de difficulté des items, Yardstick conclut que les examens modulaires du PQAP de 2016 sont légèrement plus faciles que l'ancien examen. Cette conclusion concorde avec celle qui se rapporte aux taux de réussite nationaux à l'examen. Il est possible que les taux de réussite actuels, qui sont un peu plus élevés qu'auparavant, soient attribuables au fait que les examens sont légèrement plus faciles pour les candidats. Cela dit, le niveau de difficulté de l'examen n'est qu'un des nombreux facteurs influant sur les taux de réussite à l'examen.

Lisibilité de l'examen

En plus d'avoir comparé la difficulté des items des examens du PQAP de 2015 et de 2016, Yardstick s'est penchée sur la lisibilité de l'examen. La lisibilité a été évaluée à l'aide de la formule de Flesch-Kincaid, accessible dans Microsoft Word. Cette formule repose sur l'évaluation du nombre total de phrases, de mots et de syllabes dans l'extrait d'un texte. L'application de cette formule produit un indice qui correspond à la lisibilité du texte selon le niveau scolaire du lecteur (aux États-Unis). Le nombre de phrases contenues dans le texte analysé et leur longueur sont directement proportionnels au niveau scolaire nécessaire à la lecture et à la compréhension du texte. Plus l'indice correspondant au niveau scolaire est élevé, plus l'extrait du texte est difficile à lire et à comprendre.

Les résultats de l'analyse de lisibilité de l'examen du PQAP de 2015 sont présentés au tableau 13. L'indice de lisibilité des formats d'examens du PQAP de 2015 était environ de 9.

Tableau A13. Statistiques de lisibilité de l'examen de 2015 du PQAP (formats 530, 531 et 532)

Statistiques de lisibilité	Format 530	Format 531	Format 532
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	8,7	8,7	8,8
Phrases à la voix passive (%)	5	5	5
Phrases par paragraphe	2,0	1,9	1,9
Mots par phrase	8,0	7,9	8,0
Caractères par mot	5,0	5,0	5,1

Les résultats des analyses de lisibilité portant sur les examens modulaires du PQAP de 2016 (formats 4.2, 5.2 et 6.2) sont présentés aux tableaux 14 à 17. L'indice de lisibilité des formats d'examens du module *Déontologie – Common law* se situait entre 10 et 11. L'indice de lisibilité des formats d'examens du module *Assurance-vie* se situait entre 9 et 10. L'indice de lisibilité des formats d'examens du module *Assurance contre la maladie ou les accidents* se situait entre 10 et 11. L'indice de lisibilité des formats d'examens du module *Fonds distincts* se situait entre 8 et 9. À l'exception de l'examen du module *Fonds distincts*, les indices de lisibilité des examens modulaires du PQAP de 2016 (formats 4.2, 5.2 et 6.2) étaient plus élevés d'un ou deux échelons que ceux de l'examen du PQAP de 2015.

Tableau A14. Statistiques de lisibilité de l'examen de 2016 du module Déontologie – Common law (formats 4.2, 5.2 et 6.2)

Statistiques de lisibilité	Format 4.2	Format 5.2	Format 6.2
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	10,5	10,2	10,8
Phrases à la voix passive (%)	10 %	14 %	16 %
Phrases par paragraphe	1,7	1,8	1,7
Mots par phrase	15,8	14,8	15,5
Caractères par mot	4,9	5,0	5,0

Tableau A15. Statistiques de lisibilité de l'examen de 2016 du module Assurance-vie (formats 4.2, 5.2 et 6.2)

Statistiques de lisibilité	Format 4.2	Format 5.2	Format 6.2
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	9,3	9,9	10,1
Phrases à la voix passive (%)	7 %	5 %	11 %
Phrases par paragraphe	2	2	1,8
Mots par phrase	14,2	14,8	15,2
Caractères par mot	4,8	4,9	4,8

Tableau A16. Statistiques de lisibilité de l'examen de 2016 du module Assurance contre la maladie ou les accidents (formats 4.2, 5.2 et 6.2)

Statistiques de lisibilité	Format 4.2	Format 5.2	Format 6.2
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	11,0	10,4	10,2
Phrases à la voix passive (%)	3 %	4 %	5 %
Phrases par paragraphe	2,1	2,0	2,2
Mots par phrase	13,8	13,2	12,9
Caractères par mot	5,0	4,9	4,9

Tableau A17. Statistiques de lisibilité de l'examen de 2016 du module Fonds distincts (formats 4.2, 5.2 et 6.2)

Statistiques de lisibilité	Format 4.2	Format 5.2	Format 6.2
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	8,9	9,2	9,3
Phrases à la voix passive (%)	5 %	5 %	9 %
Phrases par paragraphe	1,8	1,7	1,7
Mots par phrase	12,4	13,2	12,9
Caractères par mot	4,8	4,8	4,9

Les résultats des analyses de lisibilité portant sur les examens modulaires du PQAP de 2016 (formats 4.3, 5.3 et 6.3) sont présentés aux tableaux 18 à 21. L'indice de lisibilité des formats d'examens du module *Déontologie – Common law* se situait entre 10 et 11. L'indice de lisibilité des formats d'examens du module *Assurance-vie* se situait entre 9 et 10. L'indice de lisibilité des formats d'examens du module *Assurance contre la maladie ou les accidents* se situait entre 10 et 11. L'indice de lisibilité des formats d'examens du module *Fonds distincts* se situait entre 9 et 10. Ces indices de lisibilité sont très semblables à ceux des formats d'examens 4.2, 5.2 et 6.2.

Tableau A18. Statistiques de lisibilité de l'examen de 2016 du module Déontologie – Common law (formats 4.3, 5.3 et 6.3)

Statistiques de lisibilité	Format 4.3	Format 5.3	Format 6.3
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	10,5	10,2	10,8
Phrases à la voix passive (%)	13 %	13 %	16 %
Phrases par paragraphe	1,7	1,8	1,7
Mots par phrase	15,8	14,8	15,5
Caractères par mot	4,9	5,0	5,0

Tableau A19. Statistiques de lisibilité de l'examen de 2016 du module Assurance-vie (formats 4.3, 5.3 et 6.3)

Statistiques de lisibilité	Format 4.3	Format 5.3	Format 6.3
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	9,3	10,0	10,1
Phrases à la voix passive (%)	8 %	7 %	12 %
Phrases par paragraphe	2	1,9	1,8
Mots par phrase	14	15	15,2
Caractères par mot	4,8	4,9	4,8

Tableau A20. Statistiques de lisibilité de l'examen de 2016 du module Assurance contre la maladie ou les accidents (formats 4.3, 5.3 et 6.3)

Statistiques de lisibilité	Format 4.3	Format 5.3	Format 6.3
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	11,0	10,3	10,1
Phrases à la voix passive (%)	5 %	3 %	4 %
Phrases par paragraphe	2,2	2,0	2,2
Mots par phrase	13,9	13,5	13,0
Caractères par mot	5	4,9	4,9

Tableau A21. Statistiques de lisibilité de l'examen de 2016 du module Fonds distincts (formats 4.3, 5.3 et 6.3)

Statistiques de lisibilité	Format 4.3	Format 5.3	Format 6.3
Score de lisibilité selon le niveau scolaire de Flesch-Kincaid	9,0	9,2	9,3
Phrases à la voix passive (%)	6 %	5 %	8 %
Phrases par paragraphe	1,8	1,7	1,7
Mots par phrase	12,4	13,2	12,7
Caractères par mot	4,8	4,8	4,9

Les résultats des analyses de lisibilité tendent à indiquer que l'ancien examen du PQAP était légèrement plus facile à lire que les nouveaux examens modulaires. Il ne faut pas en conclure pour autant que les examens modulaires sont particulièrement exigeants sur le plan de la lecture (ou, en d'autres mots, qu'un candidat doit avoir une plus grande aptitude à la lecture que celle qu'on attend d'un représentant en assurance de personnes). Les nouveaux examens sont rédigés dans un niveau de langue qu'un élève de 10^{e} ou de 11^{e} année (l'équivalent de la 4^{e} ou 5^{e} secondaire) devrait pouvoir comprendre. Quelle aptitude à la lecture est attendue des représentants en assurance de personnes? La réponse à cette question contribuera à mieux évaluer la lisibilité de l'examen du PQAP.

Selon les *normes*, le niveau de langue de l'examen doit correspondre à celui que les représentants en assurance de personnes doivent maîtriser pour assumer les responsabilités liées à leur poste[2]. Les personnes qui occupent ces fonctions lisent des documents et communiquent avec des gens tous les

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

jours. Il est donc raisonnable d'affirmer que les représentants en assurance de personnes doivent avoir une aptitude à lire et à comprendre le français au moins équivalente à celle de quelqu'un qui termine ses études secondaires.

Idéalement, le niveau de maîtrise de la langue exigé dans l'exercice d'une profession doit résulter d'une analyse d'emploi spécialisée portant précisément sur les exigences linguistiques de la profession en question. Au Canada, le Centre des niveaux de compétence linguistique canadiens est un organisme qui réalise de telles analyses d'emploi. Il établit par exemple les niveaux de compétence linguistique applicables aux infirmiers, aux ergothérapeutes et aux pharmaciens. **À notre connaissance, il n'y a actuellement pas de niveau de compétence linguistique applicable aux professionnels de l'assurance. Il n'est donc pas possible de déterminer de façon formelle si le niveau de langue de l'examen du PQAP est adéquat.**

Fiabilité de l'examen

Yardstick a été invitée à commenter le lien entre la fiabilité des modules de l'examen du PQAP de 2016 et la cohérence du classement (réussite ou échec) des candidats.

Les décisions de classement des candidats peuvent être correctes ou incorrectes. Une décision de classement est jugée correcte quand le score observé et le score vrai du candidat se situent au-dessus ou en deçà de la note de passage. Une décision de classement est jugée incorrecte quand elle comporte l'une des deux erreurs de classement suivantes : un faux négatif ou un faux positif. Le *faux négatif* renvoie à un scénario où le score vrai et le score observé du candidat se situent respectivement au-dessus et en deçà de la note de passage. Le cas échéant, le candidat échoue à l'examen même s'il est suffisamment compétent pour le réussir. Le *faux positif* renvoie au scénario inverse : le score vrai et le score observé du candidat se situent respectivement en deçà et au-dessus de la note de passage. Le cas échéant, le candidat réussit l'examen même s'il n'est pas encore suffisamment compétent.

La fiabilité de l'examen a une incidence sur la justesse des décisions de classement prises à l'endroit des candidats (réussite ou échec). Quand un examen mesure le construit de façon incohérente (c.-à-d. quand l'erreur de la mesure est importante), on peut s'attendre à ce qu'un candidat obtienne un score très différent à l'examen s'il le passe de nouveau. Dans un tel contexte, il est plus probable que le score vrai d'un candidat soit très différent de son score observé (c.-à-d. qu'il se produise une erreur de classement). Toutefois, en ce qui a trait à l'examen du PQAP, les erreurs de classement ne risquent pas de nuire de façon significative à la validité de l'interprétation du score. Il y a plusieurs raisons à cela.

Premièrement, la fiabilité de trois des quatre examens modulaires du PQAP (formulaires 4.2, 5.2 et 5.3) répond aux normes conventionnelles de fiabilité (α = 0,70). La fiabilité de l'examen de 2016 du module *Déontologie – Common law* est inférieure aux normes souhaitables (α = 0,60-0,70). Cependant, il est plus probable que cette observation découle de la petite taille de l'examen que de la qualité de ses items.

L'examen de 2016 du module *Déontologie – Common law* ne contient aucun item que les candidats jugeraient excessivement difficile. En moyenne, les items donnent lieu à une réponse correcte par 70 à 72 % des candidats. Fait plus important encore, le coefficient de corrélation bisériale de point corrigé des items est élevé : seuls quelques items (n = 2-5) associés à un faible coefficient de corrélation bisériale de point corrigé ont été relevés dans chacun des formats d'examens. Enfin, le taux de réussite à l'examen de ce module en 2016 est plutôt élevé (au-delà de 80 %).

Bien que le degré de fiabilité des examens modulaires du PQAP de 2016 soit acceptable, il est toujours profitable d'évaluer comment on peut améliorer la fiabilité d'un examen. Une mesure plus précise déboucherait sur des décisions de classement erronées en diminuant la probabilité d'erreur de la mesure. En ce qui concerne l'examen du PQAP, une amélioration de la fiabilité serait possible par l'ajout d'un plus grand nombre d'items.

Deuxièmement, il n'est pas possible d'éliminer complètement les erreurs de classement dans une évaluation. Aucune mesure n'est infaillible : il y aura toujours un certain nombre de candidats dont le classement ne reflète pas le score vrai (que ce soit un faux positif ou un faux négatif). Pour réduire le risque que des décisions incorrectes soient prises à l'endroit des candidats, il serait souhaitable que les OCRA révisent les scores des candidats quand ces scores sont près de la note de passage (c.-à-d. les scores limites). Dans bon nombre d'organisations, il est monnaie courante de réviser les scores des candidats qui ont échoué de peu à un examen. Toutefois, par souci d'équité, la même pratique devrait s'appliquer aux candidats qui ont réussi l'examen de justesse. Les OCRA devraient mettre en place une procédure formelle de traitement des scores limites. Une telle procédure pourrait notamment consister à recueillir plus d'information sur les candidats ayant obtenu un score limite afin de mieux évaluer leur score vrai (ou, en d'autres mots, le score que le candidat obtiendrait à toutes les répliques possibles de la procédure de mesure).

Yardstick n'a pas effectué d'analyse formelle de la cohérence et de la justesse des décisions de classement (réussite ou échec) des candidats aux examens modulaires du PQAP de 2016, car ces décisions n'ont pas été prises d'après des notes de passage établies à l'aide de méthodes psychométriques rigoureuses et valables (p. ex., la méthode Angoff modifiée). Les indices de cohérence et de justesse des décisions fluctuent en fonction de la fiabilité et de la précision de l'examen ainsi que du niveau où se situe la note de passage. Si la note de passage n'inspire pas confiance, les indices de cohérence et de justesse des décisions seront difficiles à interpréter.

Comme elle le mentionne dans son rapport d'analyse psychométrique, Yardstick recommande aux OCRA de mener une étude formelle d'établissement de norme pour obtenir une note de passage empirique et l'erreur standard conditionnelle pour chaque module de l'examen. L'erreur standard conditionnelle servirait alors à créer un intervalle de confiance autour de la note de passage pour définir les scores maximum et minimum qui seraient obtenus pour la note de passage si la même procédure d'établissement de norme était appliquée à répétition. L'organisme de réglementation présenterait ensuite la note de passage recommandée et l'intervalle de confiance au comité consultatif sur l'examen, qui établirait le seuil de réussite de façon définitive. Le comité consultatif sur l'examen peut décider d'accepter la note de passage recommandée ou de l'ajuster à la hausse ou à la baisse en fonction des données sur l'erreur standard.

Équité de l'examen

La notion d'équité est un enjeu fondamental dans le domaine de l'administration de tests. Les organismes de réglementation doivent en tenir compte à toutes les étapes de l'élaboration et de l'administration d'un examen. Il est à noter que l'équité d'un examen « n'a pas un seul sens technique[3] ». À un haut niveau, l'Institute for Credentialing Excellence définit l'équité comme « le degré auquel les candidats à un examen sont traités de façon semblable et d'une manière raisonnable[4] ». Quand on élabore des items, il faut prendre en considération l'équité de l'examen pour veiller à ce que

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

l'évaluation du construit se fasse de façon juste entre les candidats. L'équité doit également être prise en compte dans l'administration et la notation d'un examen pour que tous les candidats passent l'examen dans les mêmes conditions et reçoivent les mêmes renseignements sur leur performance.

Les concepteurs d'examens sont formels : la meilleure façon de s'assurer de l'équité d'un examen est d'intégrer cette préoccupation aux processus d'élaboration, d'administration et de notation de l'examen. Selon la norme 3.0, toutes les étapes de l'élaboration, de l'administration et de la notation d'un examen doivent être pensées de manière à faire en sorte que les facteurs qui ne se rapportent pas à l'objectif de l'examen aient le moins de répercussions possible sur les scores de l'examen ainsi qu'à soutenir la validité de l'interprétation des scores de l'examen auprès de tous les groupes de candidats[5].

Les notions d'équité de l'examen et de validité de l'examen sont étroitement liées. Si un groupe de candidats fait l'objet d'un traitement différencié par rapport à un autre groupe de candidats, la validité de l'interprétation des scores obtenus par ces candidats à l'examen sera remise en question. Il n'est donc pas surprenant que bon nombre de mesures visant à améliorer la validité de l'interprétation des scores contribuent également à l'augmentation de l'équité de l'examen. Dans son rapport d'analyse psychométrique, Yardstick a fait un certain nombre de recommandations pour améliorer la qualité des méthodes de conception, d'administration et de notation de l'examen. Plusieurs de ces recommandations, si elles sont mises en œuvre, amélioreront *de facto* l'équité de l'examen.

Par exemple, dans son rapport d'analyse psychométrique, Yardstick évoque la nécessité de standardiser les conditions d'administration de l'examen entre les juridictions. Si les OCRA mettent en application les mêmes politiques et procédures d'administration dans l'ensemble des juridictions, tous les candidats seront traités de la même manière dans le cadre de l'administration de l'examen (équité) et leurs scores auront la même signification (validité).

Dans son rapport, Yardstick recommande aux OCRA de procéder à des examens de sensibilité du contenu de l'examen pour repérer les mots, les énoncés ou les concepts qui pourraient s'avérer offensants, stéréotypés ou plus familiers à certains groupes de candidats qu'à d'autres. Le but de cet exercice est de cibler les éléments du contenu de l'examen susceptibles d'empêcher certains groupes de candidats de mettre en valeur leur compétence. Les locuteurs non natifs du français ou de l'anglais pourraient être particulièrement désavantagés par l'utilisation d'un vocabulaire ou d'une syntaxe complexes dans la formulation des items d'un examen. Bien que le rapport ne fasse pas mention de ce groupe démographique en particulier, l'une des catégories de contenu indélicat mentionnées dans le rapport (soit « Vocabulaire peu familier à un groupe [p. ex., mots inusités ou complexes en français ou en anglais]) » s'applique à ce groupe.

L'une des recommandations de Yardstick est de recueillir et de documenter l'information sur les caractéristiques démographiques et sociodémographiques des experts de contenu ayant pris part aux études de validité, notamment celle qui sous-tend le processus d'élaboration du profil de compétences ou des spécifications de l'examen. Cette recommandation concorde avec plusieurs normes d'administration de tests citées dans le rapport, à commencer par la norme 3.3 : celle-ci précise que les sous-groupes pertinents doivent prendre part aux études de validité et de fiabilité ainsi qu'aux autres études préliminaires d'évaluation de la validité, comme le prétest.

Les experts de contenu participant à l'élaboration et à la validation des items de l'examen doivent représenter les principales caractéristiques démographiques des candidats à l'examen. Il n'est pas

[5] *Standards for Educational and Psychological Testing*, 2014, p. 63.

clairement établi si c'est le cas pour l'examen du PQAP. L'information pertinente n'était pas accessible au moment de la rédaction du rapport d'analyse de l'examen. Par conséquent, Yardstick n'a pas fait d'évaluation en bonne et due forme de la composition démographique des échantillons de l'étude de validité.

À notre connaissance, les OCRA ne recueillent pas de données démographiques sur les candidats à l'examen à l'heure actuelle. Par conséquent, il n'est pas possible pour l'instant d'évaluer l'écart moyen entre les scores obtenus à l'examen par différents groupes démographiques, y compris par les locuteurs non natifs du français ou de l'anglais. Dans la partie du rapport portant sur la traduction et l'adaptation de l'examen, Yardstick présente l'analyse du fonctionnement différentiel des items comme un moyen de reconnaître les items des examens français et anglais dont le score moyen diffère de façon statistiquement significative. Ces analyses d'items, qui pourraient s'appliquer aux locuteurs natifs et non natifs du français ou de l'anglais ou aux membres d'autres groupes démographiques ou sociodémographiques, permettraient de déterminer si la performance de certains items diffère d'un groupe à l'autre.

Yardstick reconnaît qu'il serait utile que les OCRA réalisent des analyses de sous-groupes de la performance des items et de l'examen avant et après le commencement de la période d'utilisation opérationnelle. Par exemple, si les OCRA décident de mettre en application la recommandation de Yardstick au sujet du prétest des items de l'examen du PQAP, il serait important de recueillir des données démographiques sur les participants au prétest.

Il importe toutefois de noter que selon la *norme 3.6*, le fait que la performance d'un examen ou de ses items diffère d'un sous-groupe à l'autre ne traduit pas nécessairement un manque d'équité[6]. Ces différences peuvent être observées pour des raisons légitimes : il est possible, par exemple, qu'un sous-groupe maîtrise mieux qu'un autre un aspect du contenu de l'examen. Selon la norme 3.6, des différences entre les scores obtenus à l'examen ou à certains de ses items doivent « donner lieu à des études de suivi » visant à évaluer quels aspects de la conception, du contenu ou du format de l'examen pourraient expliquer l'écart dans la performance des membres des sous-groupes.

Enfin, aucune statistique ne peut être citée pour prouver l'équité de l'examen[7]. Les OCRA devraient tenir compte de la notion d'équité et veiller à atténuer les risques d'iniquité à toutes les étapes de l'élaboration, de l'administration et de la notation de l'examen.

[6] *Standards for Educational and Psychological Testing*, 2014, p. 65.
[7] Ziecky, 2002.

Bibliographie

American Educational Research Association, American Psychological Association et 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. et Knapp, D. J. (2009). Dans Knapp, J., Anderson, L. et Wild, C. (éd.). *Certification: The ICE Handbook* (p. 67-92). Washington, DC: Institute for Credentialing Excellence.

Haladyna, T. M., Downing, S. M. et 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.

Sarrazin, G. (2003). Normes de pratique du testing en psychologie et en éducation / version française supervisée par Georges Sarrazin; édition originale produite par l'American Educational Research Association, l'American Psychological Association et le National Council on Measurement in Education, Montréal, Institut de recherches psychologiques, 231 p.

Smith, I.L. et Springer, C.C. (2009). Dans Knapp, J., Anderson, L. et Wild, C. (éd.). *Certification: The ICE Handbook* (p. 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.