

NOTE D'INFORMATION

EXAMENS D'ADMISSION AUX PROFESSIONS

Réflexions sur l'adoption d'un examen
existant et l'élaboration d'un examen

COMMISSAIRE À L'ADMISSION AUX PROFESSIONS

Mai 2025

Document approuvé le 12 mai 2025

André Gariépy, avocat, F.Adm.A., ASC
Commissaire

Rédaction par Spire Psychometrics Inc. (en anglais) :

Greg Sadesky, Ph.D.
Greg Pope, M.Sc.

Pour tout renseignement, s'adresser au bureau du commissaire :

500, boulevard René-Lévesque Ouest
6e étage, bureau 6.500, C.P. 40
Montréal (Québec) H2Z 1W7
Téléphone : 514 864-9744
Sans frais : 1 800 643-6912
Télécopieur : 514 864-9758
Courriel : commissaire@opq.gouv.qc.ca

Comment citer ce document

Commissaire à l'admission aux professions (2025). *Examens d'admission aux professions : réflexions sur l'adoption d'un examen existant et l'élaboration d'un examen*, Montréal.

Gouvernement du Québec, 2025
Tous droits réservés pour tous les pays

La reproduction de ce document est autorisée à des fins éducatives ou non commerciales à condition que l'extrait ou l'intégralité du document soit reproduit sans modification et que la source soit mentionnée.

TABLE DES MATIÈRES

INTRODUCTION.....	1
1. DÉTERMINER S'IL CONVIENT D'ADHÉRER À UNE COLLABORATION EXISTANTE EN MATIÈRE D'EXAMEN OU D'ADOPTER L'EXAMEN D'UNE TIERCE PARTIE.....	4
1.1 Sujets traités.....	6
Exemple 1. La physiothérapie.....	6
Exemple 2. L'orthophonie.....	8
1.2 Plans directeurs.....	9
1.3 Type de cadre conceptuel.....	12
1.4 Différences contextuelles.....	13
1.5 Différences réelles dans la pratique.....	14
2. MISE EN ŒUVRE D'UN EXAMEN POTENTIEL.....	15
2.1 Adoption de l'examen potentiel <i>tel quel</i>	15
2.2 Adoption de l'examen potentiel <i>avec modifications</i>	16
3. MAINTIEN ET AMÉLIORATION D'UN EXAMEN POTENTIEL APRÈS SON ADOPTION.....	20
3.1 Politique.....	20
3.2 Processus.....	20
4. ÉLABORATION D'UN EXAMEN D'ADMISSION EN INTERNE.....	22
ANNEXE A - CERTAINES NORMES SEPT APPLICABLES.....	37

LISTE DES TABLEAUX

Tableau 1 : Comparaison entre le profil de compétences canadien et les normes américaines pour l'orthophonie.....	9
Tableau 2 : Pondération des examens de physiothérapie canadiens et américains par catégories du cadre de compétences canadien	10
Tableau 3 : Plan directeur de l'ECAP en orthophonie.....	12

LISTE DES FIGURES

Figure 1 : Exemple d'analyse d'un item de la théorie classique des tests (TCT) pour un item qui répond aux normes statistiques.....	27
Figure 2 : Exemple d'analyse d'un item de la théorie classique des tests (TCT) pour un item qui ne répond pas aux normes statistiques.....	28
Figure 3 : Exemple de rapport d'analyse d'un test de la théorie classique des tests (TCT)...	28
Figure 4 : Exemple de processus d'élaboration d'un item	30

ABRÉVIATIONS, ACRONYMES ET SIGLES

ACORP :	<u>Alliance canadienne des organismes de réglementation de la physiothérapie</u>
ASHA :	<u>American Speech and Hearing Association</u> [site Web en anglais]
CC :	Cadre de compétences
CCI :	Courbes caractéristiques de l’item (Item Characteristic Curves – ICCs)
CIV	Variance non pertinente avec le construit (Construct Irrelevant Variance)
ECAP :	<u>Examens canadiens d’admission à la pratique</u>
ITC :	<u>International Test Commission</u> [site Web en anglais]
KSAs :	Connaissances, compétences et aptitudes (Knowledge, Skills and Abilities – KSAs)
NCCA :	<u>National Commission for Certifying Agencies</u> [site Web en anglais]
OAC :	<u>Orthophonie et Audiologie Canada</u>
ORD :	Organisme de réglementation destinataire
SEPT:	Normes de pratique du <i>testing</i> en psychologie et en éducation (Standards for Educational and Psychological Testing)
TRI :	Théorie de la réponse aux items (Item Response Theory – IRT)
TCT :	Théorie classique des tests (Classical Test Theory – CTT)

Note : Les hyperliens mènent au site Web de l'organisation en question.

AVANT-PROPOS DU COMMISSAIRE

La pertinence d'exiger un examen d'admission comme condition supplémentaire à une formation initiale qualifiante et pour couronner une démarche d'admission à une profession réglementée est une question qui revient de temps à autre. La question se pose tant pour les ordres professionnels qui n'ont pas ce type de condition que ceux qui en ont.

Il est toujours bon de se questionner sur la pertinence et l'utilité d'un examen. À quels besoins voulons-nous répondre par cette étape évaluative ? Est-ce le bon moyen ? Est-ce la bonne façon de faire ? Le faisons-nous en interne, seul ou dans le cadre d'une collaboration avec des organismes de réglementation de la profession de juridictions avec lesquelles nous partageons une vision similaire de la profession ? Adoptons-nous un examen existant ? Adoptons-nous un examen développé par une tierce partie et sous sa responsabilité ?

Autant de questions qui demandent des décideurs et acteurs de l'admission aux professions réglementées d'avoir une compréhension minimale des arcanes du monde des examens et de la psychométrie, discipline dont les normes et bonnes pratiques assurent la crédibilité d'outils d'évaluation.

La nature particulière et les enjeux d'un examen d'admission

Un examen d'admission est plus qu'un outil technique d'évaluation, il comprend en son sein la norme professionnelle et se construit autour d'elle. L'examen reflète aussi une vision de la profession en regard de laquelle on veut évaluer l'aptitude d'une personne candidate. Cette vision intègre, dans un tout cohérent, un propos sur :

- le contexte de la pratique ;
- le champ de pratique et les activités réservées ;
- les interactions avec les autres professionnels du domaine ;
- les compétences qui soutiennent la pratique ;
- la formation initiale qui prépare à la pratique ;
- les cadres juridique et normatif de la pratique et des services dans le domaine.

Le document

Pour répondre au besoin d'information et enrichir la réflexion sur le sujet, le Commissaire à l'admission aux professions a mandaté deux experts psychométriciens, rompus à leur discipline et à ses applications dans le contexte de l'admission aux professions réglementées, de faire part de réflexions à l'intention des ordres professionnels. Leur version du document a été rédigé en anglais. La version officielle en français est tirée d'une traduction. Le document est aussi utile aux autres acteurs de la réglementation des professions (autorités publiques, entités de surveillance, établissements d'enseignement).

Le présent document aborde le sujet en se basant sur l'hypothèse que le besoin et l'opportunité d'avoir un examen d'admission pour la profession ont été sérieusement discutés et établis. Le document traite alors en premier lieu de la question souvent soulevée d'adopter ou non un examen existant issu d'une collaboration entre organismes de réglementation de la profession de plusieurs juridictions (ex. entre des provinces et territoires canadiens) ou d'adopter un examen existant sous la responsabilité d'une tierce partie, d'ici ou d'ailleurs.

Si l'option se révèle défendable, le document présente les moyens et les défis de la mise en place d'un examen existant que l'on adopte ainsi que du suivi de la performance de celui-ci tout au long de son utilisation par l'ordre professionnel.

Enfin, le document présente les considérations entourant le développement d'un examen d'admission en interne, par l'ordre professionnel avec ou sans la collaboration d'organismes de réglementation de la profession d'autres juridictions.

Le document trace ainsi les conditions de succès et de crédibilité de l'utilisation d'un examen comme condition de délivrance d'un permis d'exercice.

Le cas d'un examen d'une tierce partie

Dans des travaux précédents, le commissaire a indiqué qu'il est possible pour les ordres professionnels de faire affaire avec une tierce partie quand celle-ci détient une expertise qui comporte des avantages pour les ordres dans l'évaluation de candidatures à l'admission. Toutefois, s'agissant d'un processus ou d'une activité qui est partie intégrante de la démarche d'admission à une profession réglementée au Québec, cela doit se faire en vérifiant certains éléments ainsi qu'en mettant en place un cadre de gouvernance et de reddition de compte adéquat.

Utiliser un examen d'une tierce partie, qui plus est hors du Québec ou du Canada, c'est le risque de perdre une partie du contrôle sur la norme professionnelle même et sur la vision de la profession. Ce n'est pas sans effet sur la formation initiale, le système professionnel, l'effectivité de la protection du public et le contexte de pratique au Québec.

La décision de recourir à un examen d'une tierce partie doit donc faire intervenir plusieurs acteurs, dont les autorités québécoises du système professionnel, ultimes garantes de l'adéquation des normes auprès du public, et les partenaires du système éducatif au Québec, qui forment les futurs professionnels.

Toutes les parties prenantes doivent se poser la question si les fondements méthodologiques de l'examen de la tierce partie de même que les normes professionnelles et la vision de la profession auxquelles il se réfère correspondent à la réalité et aux exigences de la pratique au Québec. Un examen d'une grande qualité psychométrique est peu utile s'il n'est pas aussi compatible ou en phase avec la réalité locale. Dans une démarche

d'adaptation d'un examen, il faut porter attention aux possibles biais découlant du territoire d'origine de cet examen et de la clientèle principale de candidats pour laquelle il a été conçu.

Dans le cas d'une tierce partie, a fortiori hors du Québec ou du Canada, selon la profession, si des dimensions techniques ou scientifiques peuvent être communes d'un territoire à l'autre, il peut en être autrement des dimensions non techniques, comme les aspects organisationnels, culturels et juridiques. Il en va de même pour les politiques et protocoles, le vocabulaire même. On peut aussi voir différemment l'importance relative de certains aspects et compétences dans la pratique.

Des précautions et assurances avant de s'engager envers une tierce partie

Pour le commissaire, il y a lieu pour les parties prenantes de prendre certaines précautions et s'assurer des éléments suivants avant de s'engager dans les options d'adopter un examen existant (en collaboration ou celui d'une tierce partie) ou de créer un examen en collaboration avec d'autres organismes de réglementation :

- L'examen doit être adapté à la réalité de la pratique au Québec (ex. : la grille de compétences, l'analyse de tâches et la vision de la profession doivent être compatibles ; la langue et le vocabulaire sont aussi des enjeux) ;
- L'examen doit mesurer ce que nous devons mesurer pour la pratique au Québec ;
- L'examen doit satisfaire les standards reconnus internationalement dans le domaine de la mesure et de l'évaluation, incluant ceux sur la traduction et l'adaptation de tests (contexte et culture) ;
- Le territoire qui adopte l'examen d'un autre doit être en mesure de valider le contenu pour la version à laquelle seront soumises les personnes candidates de son ressort. La seule présence symbolique et non déterminante à des comités ne suffit pas ;
- Une entente écrite devrait être conclue entre l'ordre professionnel et la tierce partie ou, selon le cas, les autres organismes de réglementation dans le cadre d'une collaboration à un programme d'examen. Celle-ci devrait prévoir une série de mesures et obligations, dont la gouvernance ainsi que l'accès au contenu de l'examen, aux analyses statistiques et aux résultats spécifiques à sa catégorie de personnes candidates, tant pour l'ordre que pour les entités gouvernementales responsables du système professionnel et de la surveillance de l'admission.

Le présent document sert à éclairer plus avant sur les aspects plus techniques qu'il convient de connaître et d'aborder pour une décision défendable et opportune dans plusieurs situations.

Mises en garde

Le point de vue exprimé dans le présent document est celui du commissaire dans l'exercice indépendant de ses fonctions prévu à la loi et celui des experts qu'il a mandaté. Ils ne s'expriment pas au nom de l'Office des professions ou du gouvernement du Québec.

INTRODUCTION¹

Il arrive souvent que les organismes de réglementation des provinces et territoires canadiens utilisent des évaluations (c'est-à-dire des tests, des examens) pour décider si les personnes candidates sont compétentes pour obtenir une certification ou une autorisation d'exercer (permis). Les évaluations que les organismes de réglementation utilisent pour autoriser ou certifier (accréditer) les individus sont habituellement les suivantes :

- Examens de connaissances académiques au seuil de compétence d'entrée à la pratique visant à confirmer qu'un candidat possède les compétences minimales dans le domaine concerné, y compris pour exercer en tant que spécialiste. En voici quelques exemples :
 - Médecine : <https://mcc.ca/fr/examens-et-evaluations/>
 - Physiothérapie : <https://alliancept.org/fr/lexamen-de-competence-en-physiotherapie-ecp/>
 - Hygiène dentaire : <https://www.fdhrc.ca/pages/ndhce/?lang=fre>
- Examens visant à confirmer que les candidats ou les inscrits (personnes déjà certifiées/licenciées) ont une bonne compréhension de leur contexte réglementaire, y compris des rôles, de l'éthique et des responsabilités de la profession. En voici quelques exemples :
 - Infirmiers auxiliaires : <https://www.clpna.com/lpn/apply-to-be-an-lpn/registration-information/jurisprudence-exam/> [en anglais]
 - Pharmacie : <https://www.ocpinfo.com/registration/registration-requirements/jp-exam/> [en anglais]
 - Technologues dentaires : <https://cdto.ca/jurisprudence-and-ethics-exams/> [en anglais]
- Examens qui mènent à des certifications qui permettent aux personnes titulaires d'une autorisation d'exercer d'élargir leur champ d'action ou de faire progresser leur carrière. En voici quelques exemples :
 - Technologues en ingénierie : <https://www.aset.ab.ca/Home/Registration/Resources/ASET-Certification-Exam> [en anglais]
 - Technologues en radiologie : <https://www.camrt.ca/fr/agrement-2/programmes-detudes-canadiens-agrees/examen-dagrement-dentree-a-la-pratique/inscription-a-lexamen-dagrement-dentree-a-la-pratique-de-lactrm/>
 - Gestion de l'information sanitaire : <https://cchim.ca/fr/certification/national-certification-examinations/>

¹ Auteurs à compter de cette section : Greg Sadesky, Ph.D. et Greg Pope, M.Sc., Spire Psychometrics Inc.

L'approche la plus courante consiste, pour les organismes de réglementation, à développer leurs propres évaluations « en interne » (par ou au sein de la communauté de la profession), individuellement ou en collaboration (par exemple, en créant une entité ou un programme d'examen pancanadien). Cette approche présente des avantages, notamment

- Les délais d'élaboration et de mise en œuvre de l'examen interne sont contrôlés par une organisation de la profession, sous la responsabilité des organismes de réglementation.
- Adaptation de l'examen spécifique au contexte de la pratique et au groupe de candidats visé.
- Amélioration de l'adhésion des parties prenantes étant donné que l'examen serait élaboré par des experts en la matière au sein de la juridiction (province ou territoire) ou des différentes juridictions participantes.
- Les exigences linguistiques (c'est-à-dire la traduction des questions) seraient plus adaptées aux candidats cibles au sein de la juridiction ou des différentes juridictions participantes.
- Réduction du risque que l'examen ne réponde pas à toutes les exigences de la profession et de sa communauté, en particulier des organismes de réglementation.
- La propriété du programme d'examen signifie que son avenir est sous le contrôle de la profession et de sa communauté, en particulier des organismes de réglementation, et non d'une organisation tierce éloignée (par exemple, si une tierce partie décide de cesser d'administrer l'examen ou de modifier l'examen de manière significative).
- Les coûts sont sous le contrôle de la profession et de sa communauté, en particulier des organismes de réglementation, de même que les recettes du programme d'examen. De nombreux programmes d'examens internes sont rentables au bout d'un certain temps (deux ans, par exemple).

Malgré ces avantages, le développement en interne d'une évaluation psychométrique défendable à enjeux élevés présente des difficultés. Il s'agit notamment :

- Des coûts associés à la rémunération des experts en la matière pour leur temps, le personnel interne pour gérer le programme d'examen, le développement d'items et la banque afférente, les services psychométriques et la livraison de l'examen. Les petits organismes de réglementation, voire les groupes d'organismes de réglementation, peuvent ne pas disposer des ressources nécessaires pour poursuivre eux-mêmes le développement d'un programme d'évaluation.

- De la limitation des ressources humaines pour les petites organisations (par exemple, il se peut que le personnel interne ne soit pas en mesure d'assurer la surveillance d'un programme d'examen en cours d'élaboration).
- De la disponibilité d'experts psychométriques pour travailler sur le programme d'examens internes
- Du risque accru pour l'organisation, les candidats et les parties prenantes si le programme d'examen interne ne répond pas aux normes psychométriques applicables aux tests à enjeux élevés.

Certains organismes de réglementation des provinces et territoires canadiens expriment le besoin d'examens valables sur le plan psychométrique pour déterminer si les personnes cherchant à se faire accréditer dans leur juridiction sont compétentes. Les organismes de réglementation ont alors le choix entre élaborer un examen d'accréditation en interne (individuellement ou en collaboration avec d'autres organismes de réglementation provinciaux) ou faire appel à un programme d'examen d'une tierce partie.

L'objectif de ce document est de fournir des réflexions sur une méthodologie psychométriquement défendable pour les organismes de réglementation qui cherchent à se joindre à une collaboration existante au sein de la profession autour d'un programme d'examen ou à adopter un programme d'examen d'une tierce partie dans leur juridiction afin d'évaluer la défendabilité de cet examen. Ce document fournit des réflexions également sur la manière de mettre en œuvre et de maintenir un programme d'examen au sein de la profession ou un examen d'une tierce partie lorsqu'un tel programme a été choisi.

1. DÉTERMINER S'IL CONVIENT D'ADHÉRER À UNE COLLABORATION EXISTANTE EN MATIÈRE D'EXAMEN OU D'ADOPTER L'EXAMEN D'UNE TIERCE PARTIE

Lorsqu'un organisme de réglementation envisage de s'associer à une collaboration existante en matière d'examens au sein de la profession ou d'utiliser un examen d'une tierce partie pour son propre programme de certification ou d'autorisation d'exercer, il y a en principe trois résultats possibles :

- 1) L'organisme de réglementation peut adopter l'examen « tel quel » ;
- 2) L'organisme de réglementation peut adopter l'examen, mais en l'adaptant de manière à ce qu'il corresponde le mieux possible à son contexte et à ses besoins ;
- 3) L'organisme de réglementation peut rejeter l'examen et envisager d'autres options pour les examens, y compris l'élaboration d'un examen interne, individuellement ou dans le cadre d'une nouvelle collaboration avec des organismes provinciaux de réglementation.

Le caractère souhaitable de chacun de ces résultats est déterminé par une combinaison de l'adéquation de l'examen potentiel et du contexte qui a conduit l'organisme de réglementation à envisager un examen en premier lieu. Naturellement, des options moins objectivement souhaitables peuvent être adoptées lorsqu'il existe peu d'autres solutions viables. Aux fins du présent document, il est supposé qu'un organisme de réglementation exprime le besoin d'un examen pour déterminer l'aptitude à la pratique de ses candidats. Il n'existerait pas d'alternative viable à un examen à cette fin.

À la lumière de l'hypothèse ci-dessus, le processus décisionnel qui aboutira à l'un des résultats susmentionnés repose presque entièrement sur la **similitude** entre le contenu sur lequel l'examen potentiel est basé et le contenu que l'organisme de réglementation souhaite évaluer. En outre, plusieurs aspects de cette similitude doivent être évalués, à savoir

- 1) Sujets traités
- 2) Les poids relatifs attribués aux sujets au sein du plan directeur de l'examen
- 3) Les cadres conceptuels qui sous-tendent le contenu de l'examen (par exemple, le profil de compétences)
- 4) Similitudes et préférences linguistiques
- 5) Spécificités régionales

La première facette de la similitude, qui est de loin la plus importante, est la matière décrite dans chaque cadre de compétences (CC). Si l'on peut s'attendre à ce que les examens de certaines professions portent sur des sujets similaires, indépendamment de la juridiction, il n'en va pas de même pour d'autres professions. Cela peut se produire lorsque le champ d'exercice, les normes de pratique, l'environnement technologique et l'organisation du travail dans le secteur diffèrent d'une juridiction à l'autre. Par exemple, une profession de santé spécialisée telle que la physiothérapie, l'hygiène dentaire ou l'orthophonie devrait être plus similaire qu'une profession dont le domaine est moins clairement défini, telle que celle de *responsable de la sécurité ou d'aide-soignant*. Cette conclusion est raisonnable pour de nombreux observateurs, car la base de ces professions de santé spécialisées, le corps humain, est fondamentalement la même d'un pays à l'autre et, par conséquent, les compétences et les connaissances qui les sous-tendent sont susceptibles d'être similaires. Cependant, l'inclusion d'autres facteurs tels que le contexte et les compétences dites non techniques dans les sujets d'un examen peut diminuer la similitude entre le contenu des deux programmes de certification.

Même lorsque les sujets sont similaires, d'autres différences peuvent affecter l'applicabilité de l'examen potentiel dans le contexte de l'organisme de réglementation. L'une de ces différences potentielles concerne l'importance relative du contenu, tel qu'elle ressort des plans directeurs d'examen. Bien que la matière globale puisse être très similaire, si les poids des sujets au sein des plans directeurs d'examen sont significativement différents, les conclusions sur la compétence tirées de l'examen peuvent différer de l'intention de l'organisme de réglementation destinataire (ORD). Par exemple, imaginons une profession de santé hypothétique pour laquelle l'entité responsable de l'examen potentiel et l'organisme de réglementation destinataire ont tous deux défini la catégorie de compétence « communication ». Imaginons également que, pour l'ORD, cette catégorie ait un poids deux fois supérieur à celui de l'examen potentiel. Dans ce cas, une performance élevée en communication profitera davantage aux candidats de l'ORD qu'à ceux de l'examen potentiel, ce qui pourrait modifier les conclusions sur les compétences d'un candidat et limiter l'applicabilité de l'examen potentiel dans le contexte de l'ORD. En raison de ces effets potentiels, les différences de pondération des plans directeurs doivent être réconciliées avant qu'un examen potentiel puisse être adopté.

Une autre source potentielle de différence est le cadre conceptuel sur lequel repose l'examen potentiel et le cadre conceptuel correspondant pour l'organisme de réglementation qui envisage de l'adopter. Au niveau le plus fondamental, un examen génère des preuves de la possession d'éléments du cadre conceptuel auquel il est lié. Lorsque deux cadres conceptuels sont substantiellement similaires, la preuve de la possession d'éléments de l'un peut être interprétée comme une preuve de la possession d'éléments de l'autre. Pour que cette approche fonctionne, il est essentiel de démontrer que les cadres conceptuels sont suffisamment similaires pour fonctionner de manière presque interchangeable. Comme nous le verrons dans les sections suivantes, ce seuil est rarement possible sans l'adoption complète du CC de l'examen potentiel, ou ne peut être atteint que pour une partie limitée du CC.

La langue et d'autres différences régionales peuvent également influencer sur l'applicabilité d'un examen à l'ORD. Un exemple simple concerne les différents systèmes de mesure en vigueur aux États-Unis et au Canada. Si un ORD canadien souhaite adopter un examen américain et que toutes les questions portant sur des quantités mesurées sont formulées en unités impériales, le candidat canadien sera nettement désavantagé, même s'il tente de se familiariser au préalable avec le système impérial. Les différences régionales peuvent également inclure la prévalence accrue de certains troubles dans des régions géographiques particulières. Si un examen potentiel contient des références à des maladies spécifiques à une région, et donc moins pertinentes, tout en omettant les maladies les plus répandues dans la région de l'ORD, la performance à l'examen ne reflétera pas clairement la compétence telle qu'elle est entendue par l'ORD.

Dans la section suivante, chacune de ces facettes et les résultats qui découlent de leur évaluation seront décrits plus en détail.

1.1 Sujets traités

Il est peut-être évident que la mesure dans laquelle un examen potentiel peut être adopté sans modification est fortement déterminée par la mesure dans laquelle les sujets de l'examen sont similaires à la définition de ces sujets pour l'ORD. À première vue, il peut sembler que certaines professions soient susceptibles d'être définies de manière similaire, indépendamment de la région ou de la juridiction. Les professions de la santé, telles que la physiothérapie, l'hygiène dentaire et l'orthophonie, semblent être des exemples probables de cette similitude, étant donné que ce qui sous-tend ces professions, le corps humain, est substantiellement similaire au niveau du groupe, d'un point de vue axiomatique.

Malgré la similitude « de principe » et donc l'attente que les examens d'un contexte soient utilisables dans un autre, il y a un certain nombre de raisons pour lesquelles ce ne pourrait pas être le cas. Tout d'abord, bien qu'une profession dans plusieurs juridictions puisse se référer à la même physiologie ou au même domaine sous-jacent, l'inclusion de connaissances et de compétences *non techniques* dans la définition de la profession peut affecter la comparabilité. En particulier, lorsque la définition de la profession d'une région inclut des compétences complémentaires telles que le contexte, la communication, les responsabilités professionnelles et la collaboration, et que l'autre ne le fait pas, les résultats d'un examen élaboré pour une région ne peuvent pas être utilisés sans modification dans l'autre région.

Exemple 1. La physiothérapie

Prenons l'exemple de la physiothérapie au Canada par rapport aux États-Unis. Au Canada, la définition de la profession est fournie par le [profil de compétences](#) établi en 2017 par l'Alliance canadienne des organismes de réglementation de la physiothérapie ([ACORP](#)). Le profil est structuré en 7 domaines et décrit à la fois les *compétences essentielles*, définies

comme « une capacité requise par un physiothérapeute », et les étapes d'accès à la pratique, une capacité liée à la compétence essentielle attendue pour l'accès à la pratique. Les domaines sont les suivants

1. Expertise en physiothérapie
2. La communication
3. La collaboration
4. La gestion
5. Le leadership
6. L'érudition (recherche)
7. Le professionnalisme

Aux États-Unis, les sujets de l'examen ne sont pas explicitement liés à un cadre de compétences. Bien que plusieurs exemples puissent être trouvés (exemples en anglais, [ici](#) et [ici](#)), l'examen national n'y fait pas explicitement référence. Au lieu de cela, l'examen fait référence à une *analyse des tâches professionnelles*, ou à un inventaire des tâches professionnellement pertinentes effectuées par les physiothérapeutes en exercice dans le cadre de leur travail. Cette analyse est régulièrement mise à jour, mais s'apparente davantage à une vaste enquête sur la prévalence des tâches effectuées qu'à des exigences de *principe* pour une performance compétente.

En raison des différences de processus dans la conception et la création de l'examen, l'information la plus fiable pour comparer les similitudes est peut-être le contenu présenté dans le [plan directeur](#) de l'examen [en anglais]. Ce document montre que le contenu de l'examen et, par conséquent, les compétences en physiothérapie sont divisées de la manière suivante :

1. Système corporel
 - a. Systèmes cardiovasculaire et pulmonaire
 - b. Système musculo-squelettique
 - c. Systèmes neuromusculaire et nerveux
 - d. Système tégumentaire
 - e. Systèmes métabolique et endocrinien
 - f. Système gastro-intestinal
 - g. Système génito-urinaire
 - h. Système lymphatique
 - i. Interactions des systèmes
2. Hors système
 - a. Équipements, dispositifs et technologies
 - b. Modalités thérapeutiques
 - c. Sécurité et protection
 - d. Responsabilités professionnelles
 - e. Recherche et pratique fondée sur des données probantes

Il n'est pas facile de comparer ces deux descriptions de haut niveau du contenu des examens, car les cadres de description des compétences et des connaissances requises divergent considérablement. Ce que l'on peut conclure, c'est que les chevauchements entre les contenus des examens des États-Unis et du Canada diffèrent de manière significative. Les exemples les plus évidents de cette différence comprennent l'omission totale des catégories Communication et Collaboration dans le plan directeur américain et le manque d'informations spécifiques dans le cadre canadien concernant les systèmes corporels à évaluer. Comme nous le verrons dans la section suivante, des informations supplémentaires sur les systèmes se trouvent dans le plan directeur canadien, mais ces informations démontrent également de manière plus aiguë les dissemblances entre les deux examens.

Exemple 2. L'orthophonie

L'orthophonie est un exemple de profession dont les cadres de compétences et d'examens américains et canadiens sont plus similaires. Comme la physiothérapie, les domaines du [profil de compétences](#) sont structurés par 7 rôles², et décrivent les compétences dites *essentiels*, définies comme « les capacités que chaque clinicien doit être en mesure de démontrer lors de son entrée dans la pratique au Canada ». Les rôles sont les suivants :

1. Expert (connaissances, clinique)
2. Communicateur
3. Collaborateur
4. Défenseur
5. Chercheur
6. Gestionnaire
7. Professionnel

Le cadre américain correspondant a été créé par l'American Speech and Hearing Association (ASHA) à partir d'une analyse de la pratique et du programme d'études de l'orthophonie. Ces normes sont présentées ci-dessous. Étant donné que seules certaines de ces normes décrivent les compétences requises à l'entrée dans la pratique, seul le sous-ensemble correspondant des normes est pertinent pour la comparaison des cadres de compétences. Les trois normes pertinentes pour la comparaison sont indiquées en gras.

- Norme I- Degré
- Norme II - Programme d'éducation
- Norme III- Programme d'études
- **Norme IV - Résultats en matière de connaissances**
- **Norme V - Résultats en matière de compétences**

² Les profils de compétences en physiothérapie et en orthophonie ont été fortement influencés par la structure du profil [CanMeds](#) (2015), qui organise la pratique professionnelle compétente autour des multiples rôles du médecin dans la prestation des soins.

- Norme VI - Évaluation
- **Norme VII - Fellowship clinique en orthophonie**
- Norme VIII - Maintien de la certification

Le Tableau 1 ci-dessous montre la correspondance entre les cadres canadien et américain. Comme on peut le constater, des aspects de chacun des sept rôles sont présents dans les normes, en particulier dans les rôles d'*expert*, mais aussi fortement représentés dans les rôles de communicateur, de collaborateur et de professionnel. En raison de cette similitude, l'utilisation d'un examen dérivé des normes américaines pourrait convenir au Canada. Une évaluation plus approfondie des similitudes à l'aide des plans directeurs de chaque pays pour la physiothérapie et l'orthophonie sera présentée dans la section suivante.

Tableau 1 : Comparaison entre le profil de compétences canadien et les normes américaines pour l'orthophonie

		Competency Profile Roles (Canada)						
		Expert	Communicator	Collaborator	Advocate	Scholar	Manager	Professional
Standards (US)	IV - A	2						
	IV - B	2						
	IV - C	2						
	IV - E							6
	IV - F	2				1		
	IV - G	6	6	2				3
	IV - H		1	2	5	3	1	3
	V - A		10	6				
	V - B	33		6				
	V - F	1						
	VII - A						6	
	VII - C							12

1.2 Plans directeurs

Comme nous l'avons montré ci-dessus en examinant les sujets, l'adéquation d'un examen potentiel peut être provisoirement évaluée en déterminant la similitude entre la définition du contenu de cet examen et la définition correspondante de l'ORD. Une évaluation plus précise de la similitude peut être effectuée en analysant les plans directeurs pour chacun d'entre eux. Ces plans fournissent non seulement la structure catégorielle et conceptuelle sous-jacente à chaque examen, mais aussi le poids relatif de chacune des catégories et des éléments (par exemple, les compétences) qui les composent.

En comparant les plans directeurs de physiothérapie américain et canadien, les différences entre les examens deviennent évidentes. Si l'on considère les plans directeurs du point de vue de la pondération relative des compétences techniques par rapport aux compétences non techniques, comme le montre le Tableau 2 ci-dessous, l'examen canadien met beaucoup plus l'accent sur les compétences non techniques que son homologue américain. En supposant que l'*expertise en physiothérapie* représente la compétence

technique pour l'examen canadien et le contenu du *système corporel* pour l'examen américain, l'examen canadien a environ **30 % moins** de contenu en compétence technique que l'examen américain.

Tableau 2 : Pondération des examens de physiothérapie canadiens et américains par catégories du cadre de compétences canadien

	Catégorie du CC canadien	Plan directeur canadien	Plan directeur américain	Différence approx.
Technique	1. Expertise en physiothérapie	55 %	85 %+	-30 %
Non technique	2. Communication	10 %	N/A	+10 %
	3. Collaboration	10 %	N/A	+10 %
	4. Gestion	5 %	5 %	0 %
	5. Leadership	5 %	N/A	+5 %
	6. Érudition (recherche)	5 %	N/A	+5 %
	7. Professionnalisme	10 %	<5 %	+5 %+

Note : La mention « N/A » indique que certains aspects de la catégorie peuvent être représentés dans le contenu du test, mais pas explicitement dans les documents de référence respectifs. Par conséquent, la somme des pondérations américaines n'est pas égale à 100 %.

Les différences considérables entre les plans directeurs présentés dans le tableau ci-dessus ont des conséquences importantes sur l'utilisation potentielle de l'examen américain dans le contexte canadien. Étant donné que chaque plan directeur représente le jugement de chaque pays sur l'importance de la catégorie de compétences respective dans la détermination de la compétence d'accès à la pratique, **et donc de l'octroi potentiel d'une autorisation d'exercer**, la performance à un examen n'est qu'une preuve limitée de la compétence définie par la juridiction. En d'autres termes, une performance élevée à l'examen de physiothérapie américain, très technique, n'implique même pas une compétence de niveau débutant par rapport à des normes moins techniques, comme le montre le plan directeur canadien. Dans ce contexte, l'examen américain ne pourrait servir de preuve que pour une partie de l'ensemble des compétences à évaluer pour l'obtention de l'autorisation d'exercer au sein du Canada. Cependant, même cette utilisation limitée de l'examen américain est suspecte ; les proportions relatives des systèmes corporels représentés dans l'examen américain sont très différentes de celles de l'examen canadien. Par exemple, pour les questions basées sur les systèmes, le pourcentage de contenu sur le système musculo-squelettique est désigné comme étant de 30 % pour l'examen américain et de 50 % pour l'examen canadien. Par conséquent, l'utilisation de la seule partie technique de l'examen américain comme preuve de la possession des compétences techniques correspondantes par rapport aux normes canadiennes n'est pas clairement justifiée.

Cet exemple montre clairement que l'utilisation de l'examen américain de physiothérapie n'est non seulement pas simple, mais qu'elle créerait des risques spécifiques de classification erronée des candidats canadiens en raison de la représentation limitée des compétences non techniques.

Il convient de noter qu'aucune des analyses ci-dessus ne suggère de manière définitive que les performances des **aspects techniques** de la profession diffèrent d'une région à l'autre. La prévalence sous-jacente de certaines affections peut varier pour les raisons évoquées ci-dessus, mais l'évaluation, le diagnostic et le traitement des mêmes troubles ne varieront probablement pas plus qu'entre les praticiens d'une même région. C'est plutôt au niveau des compétences auxiliaires et des contextes réglementaires et culturels que l'exercice d'une même profession devrait différer. Étant donné que ces compétences et contextes font partie du contenu de l'examen pour de nombreuses professions, l'efficacité d'un examen potentiel pour distinguer les candidats qui sont compétents au moment de l'entrée dans la profession de ceux qui ne le sont pas sera diminuée par l'ampleur de ces différences.

En ce qui concerne l'orthophonie, on constate une meilleure correspondance entre les examens potentiels, mais certains problèmes d'équivalence subsistent. Le [plan directeur](#) [en anglais] de l'examen de l'American Speech Language Hearing Association (ASHA) révèle une répartition égale entre trois catégories :

- *Fondements et pratique professionnelle,*
- *Dépistage, évaluation et diagnostic, et*
- *Planification, mise en œuvre et évaluation du traitement.*

À titre de comparaison, les examens canadiens d'admission à la pratique (ECAP) supervisés par Orthophonie et Audiologie Canada (OAC) spécifient le plan directeur détaillé présenté dans le Tableau 3. Les catégories du plan directeur Praxis (ASHA) sont larges ; de nombreux sujets différents peuvent être couverts dans chacune d'entre elles. En outre, bien que l'ASHA décrive cette diversité au sein de chaque catégorie sur son site Web, la répartition en pourcentage de chaque sujet au sein de chaque catégorie n'est pas disponible. Il est donc impossible d'établir une comparaison détaillée entre les deux plans directeurs. Cela dit, la comparaison relative entre le contenu technique et non technique semble plus similaire pour l'orthophonie que pour la physiothérapie. Plus précisément, le pourcentage de contenu technique par rapport au contenu non technique semble largement comparable. Pour l'examen ECAP, le pourcentage de contenu technique, c'est-à-dire lié à un rôle d'expert, est en moyenne de 65 %. Si la première catégorie de l'examen de l'ASHA en orthophonie est considérée comme non technique et que chacune des trois catégories a la même pondération, le contenu technique correspondant serait de 67 %.

Un bémol à la conclusion ci-dessus est que la composante « Fondements » de la première catégorie peut être considérée au moins comme un précurseur des connaissances techniques, car elle concerne le développement selon la durée de vie, les facteurs influençant la communication et l'épidémiologie. Plus ce contenu est important dans la

première catégorie, moins les deux examens seront similaires. Bien entendu, moins il y a de similitudes entre les examens possibles, moins l'examen américain sera applicable au contexte canadien. Il convient de mentionner qu'au moins une [juridiction](#) canadienne [site Web en anglais] a jugé l'examen Praxis de l'orthophonie suffisamment similaire pour l'adopter afin d'éclairer les décisions relatives à l'autorisation d'exercer. La décision d'adopter cet examen pourrait être justifiée par rapport à d'autres options plus exigeantes en ressources, telles que l'élaboration d'un tel examen en interne de manière individuelle. Cependant, il ne fait aucun doute que tout examen développé en interne, en supposant que les meilleures pratiques soient respectées, sera mieux adapté à son propre contexte qu'un examen d'une tierce partie extérieur à la juridiction.

Tableau 3 : Plan directeur de l'ECAP en orthophonie.

Contenu de l'ECAP en orthophonie Catégories	Nombre de questions	Pourcentage de l'examen
1. Rôle d'expert - Connaissances	34 - 51	20 - 30 %
2. Rôle d'expert - Clinique	60 - 77	35 - 45 %
3. Rôle du communicateur	9 - 25	5 - 15 %
4. Rôle du collaborateur	5 - 12	3 - 7 %
5. Rôle de défenseur	5 - 12	3 - 7 %
6. Rôle du gestionnaire	5 - 12	3 - 7 %
7. Rôle professionnel	9 - 25	5 - 15 %

Les deux exemples décrits ci-dessus fournissent une base concrète pour anticiper le degré d'applicabilité d'un examen potentiel pour un organisme de réglementation provincial et territorial canadien. L'élément le plus important pour déterminer cette applicabilité est la mesure dans laquelle la matière (les sujets) et les plans directeurs sont similaires entre l'examen potentiel et celui de l'ORD. Ces exemples montrent en particulier que la mesure dans laquelle les compétences et les connaissances non techniques sont représentées dans les spécifications du contenu de chaque examen est un facteur limitatif important de l'applicabilité. D'autres facteurs moins fondamentaux sont décrits ci-dessous.

1.3 Type de cadre conceptuel

Dans la section ci-dessus, le contenu au niveau de granularité le plus élevé a été examiné pour déterminer les similitudes générales. Des descriptions plus spécifiques des connaissances et des compétences sont également importantes, à la fois en tant que facteur conduisant à certaines de ces différences générales et dans la manière dont ce contenu est défini. On peut soutenir que l'une des raisons pour lesquelles les examens canadiens décrits et comparés sont composés d'un contenu plus « non technique »

concerne le processus de définition des compétences et des connaissances spécifiques à la profession. Au Canada, le cadre conceptuel qui sous-tend les évaluations de certification et d'autorisation d'exercer est généralement basé sur les compétences, ce qui signifie que les professions sont définies en termes de connaissances et d'aptitudes sous-jacentes nécessaires pour les exercer avec succès. En revanche, aux États-Unis, la méthode dominante de description d'une profession à des fins d'évaluation consiste à décrire des actions ou des fonctions spécifiques que le professionnel accomplit, appelées « tâches professionnelles ». Par nature, le développement des compétences est un exercice de conceptualisation de ce qui permet d'accomplir avec succès une tâche professionnelle. L'analyse des tâches professionnelles est plutôt un inventaire des activités spécifiques à la profession que les praticiens performants accomplissent déjà. Prenons l'exemple suivant, qui décrit la capacité d'un praticien à remplir des fonctions d'équipe :

Compétence : *Faciliter la communication au sein de l'équipe*

Mission : *Conduire des réunions d'équipe*

Cet exemple permet de mieux comprendre les différences entre les compétences et les tâches professionnelles. La compétence décrit l'attribut qu'un candidat retenu doit posséder pour exercer efficacement sa profession, tandis que la tâche professionnelle décrit l'activité spécifique liée à l'emploi. L'importance de cette différence réside dans le fait que les questions d'évaluation qui sont ostensiblement axées sur la même compétence peuvent être créées de manière très différente en raison de la manière dont cette compétence sous-jacente est décrite. Une question axée sur la compétence est plus susceptible de cibler les moyens de favoriser une communication efficace, tandis qu'une question axée sur les tâches professionnelles est plus susceptible de faire référence aux caractéristiques des réunions efficaces. Sans débattre des mérites d'une approche par rapport à l'autre, le contenu sous-jacent de l'examen est susceptible de s'aligner sur la manière dont les compétences et les connaissances sont décrites. Par conséquent, des différences apparaîtront dans la manière dont les candidats sont évalués, ce qui limitera encore davantage les similitudes.

1.4 Différences contextuelles

Comme nous l'avons mentionné dans l'introduction de cette section, des différences telles que les systèmes de mesure peuvent entraîner des difficultés sélectives pour les candidats dans certaines juridictions par rapport à d'autres. Cette sélectivité érode la similitude. Outre les unités de mesure, il existe de nombreuses autres sources de différences organisationnelles, linguistiques et culturelles. Le champ de pratique peut être différent. De même, certaines professions désignent leurs clients par le terme « clients » ou « patients » ; des juridictions différentes peuvent également utiliser des termes différents. En outre, ces différences peuvent s'étendre à tous les aspects de la pratique. Plus ces différences existent et sont rencontrées par les candidats à l'examen potentiel, plus l'examen sera difficile.

Le contenu potentiel de l'examen qui n'est pas pertinent pour la juridiction de l'ORD est potentiellement plus nuisible. Par exemple, sous la rubrique « Responsabilités professionnelles » du plan directeur de la physiothérapie aux États-Unis se trouve le point suivant :

- Normes de facturation, de codage et de remboursement³.

Il est bien connu que le système d'assurance maladie est très différent entre les deux pays et, par conséquent, un Canadien qui passe l'examen de physiothérapie n'est probablement pas au courant des normes susmentionnées. Pire encore, ces normes ne sont pas pertinentes pour la pratique professionnelle au Canada. Si des ressources étaient mises à la disposition des candidats canadiens pour qu'ils puissent répondre correctement à ces questions, leurs notes pourraient être gonflées et leur compétence par rapport aux normes canadiennes mal évaluée. Dans le cas extrême, cette petite différence pourrait conduire à ce qu'un candidat soit jugé comme ayant réussi l'examen alors que ses capacités réelles ne justifieraient pas cette conclusion ou l'autorisation d'exercer associée.

1.5 Différences réelles dans la pratique

Même lorsque le champ d'exercice et les normes d'exercice sont similaires d'une juridiction à l'autre, l'exercice de la profession peut différer. Par exemple, la prévalence de certaines conditions et les données démographiques peuvent modifier la nature de la pratique et donc influencer le contenu de l'examen. Par exemple, on sait que les taux d'obésité [sont plus élevés](#) [site Web en anglais] aux États-Unis que dans d'autres pays, et on peut donc raisonnablement s'attendre à ce que l'examen d'une profession connue pour traiter des problèmes liés à l'obésité ait plus de contenu lié à ces problèmes. En revanche, un examen conçu principalement pour une autre juridiction pourrait accorder moins d'importance à des pathologies plus répandues au Canada.

³ On peut se demander si ces éléments non cliniques représentent des exigences d'accès à la pratique et s'ils doivent donc faire partie des compétences de la profession et du plan directeur de l'examen. Il s'agit davantage de questions administratives qui sont généralement apprises dans un cadre différent lors de l'entrée dans la profession ou dans l'emploi.

2. MISE EN ŒUVRE D'UN EXAMEN POTENTIEL

La section ci-dessus décrit les critères généraux permettant de déterminer s'il est approprié d'adhérer à une collaboration existante en matière d'examen au sein de la profession ou d'adopter l'examen d'une tierce partie pour l'utiliser dans le cadre d'un ORD. La base de cette détermination est la similitude du contenu de l'examen tel que défini principalement par les sujets et le plan directeur, mais aussi par d'autres critères tels que la structure des compétences, les différences contextuelles et les différences réelles dans la pratique. Lorsque l'examen potentiel est jugé suffisamment similaire, l'ORD peut décider d'adopter l'examen *tel quel*, dans son intégralité, en adoptant tous ses processus et résultats. Sauf s'il n'y a pas d'autres solutions et/ou ressources, comme la collaboration entre organismes de réglementation, ce résultat devrait être rare; il y aura peu d'exceptions où la concordance entre un examen potentiel et les besoins d'un ORD se chevauchera à ce point. Il est plus probable que l'ORD adopte l'examen potentiel avec des modifications spécifiques ou en partie.

Cette section aborde à la fois les situations dans lesquelles un ORD peut choisir d'adopter un examen potentiel dans son ensemble et les types de modifications qui permettent à un examen potentiel d'être utilisé d'une manière qui correspond mieux aux besoins et aux contraintes de l'ORD.

2.1 Adoption de l'examen potentiel tel quel

Comme mentionné ci-dessus, il existe relativement peu de situations dans lesquelles l'adoption complète d'un examen potentiel pourrait être considérée comme justifiée d'un point de vue psychométrique. L'alignement entre les définitions, principalement les sujets et le plan directeur, d'une profession, d'une juridiction à l'autre, sera suffisamment différent pour que la performance sur un contenu non pertinent ou mal aligné conduise à des scores et donc à des conclusions qui ne sont pas suffisamment précis pour leur objectif. Dans les rares cas où cette démarche est justifiée, le profil de compétences (ou l'analyse des tâches professionnelles) devrait être harmonisé entre les juridictions (par exemple, dans les différents organismes de réglementation provinciaux et territoriaux canadiens qui se réfèrent au même profil) ou l'ORD devrait avoir effectué une analyse exhaustive du contenu de l'examen par des experts en la matière, en vérifiant que les éléments sont pertinents et que, pris ensemble, ils représentent proportionnellement bien les compétences et les connaissances requises.

Une autre option pour l'adoption complète est possible, à savoir que l'ORD décide non seulement d'adopter l'examen potentiel, mais aussi les sujets et le plan directeur comme étant les siens. Cette voie entraîne des conséquences importantes et difficiles, car les programmes d'enseignement sont probablement déjà orientés vers la préparation des étudiants à la profession telle qu'elle a été définie précédemment. Par conséquent, au moins pendant les premières années, les résultats à l'examen sont susceptibles de chuter

et de se redresser au fur et à mesure que l'écart entre la matière représentée à l'examen et celle enseignée dans les programmes de formation se réduit. En outre, il y a probablement d'excellentes raisons pour lesquelles une juridiction a décidé d'une définition particulière de la profession, représentée par un cadre réglementaire, des programmes de formation agréés ou reconnus, un profil de compétences ou d'autres CC. L'adoption de l'examen potentiel pourrait ne pas les prendre en compte de manière adéquate.

La qualité de la définition de la profession et du cadre réglementaire fondée sur les CC est au cœur de la décision d'adopter un examen potentiel. La protection du public est étroitement liée à une pratique professionnelle compétente et la compétence elle-même est définie par le CC. Bien qu'une analyse complète de la concordance d'un CC avec les normes d'exercice dépasse le cadre du présent document, il est essentiel de reconnaître que l'adoption complète du CC simplement pour utiliser un examen potentiel revient à rejeter des normes spécifiques à une juridiction en faveur de celles d'une autre. Bien sûr, les normes peuvent être suffisamment similaires pour que ce compromis soit acceptable, mais cette décision doit être prise en étant pleinement conscient de ce que l'on abandonne, à savoir le contrôle de la définition des normes professionnelles à l'entrée dans la pratique, qui est souvent une prérogative des autorités publiques provinciales et territoriales au Canada ou qui doivent au moins être informées de la démarche.

Il convient également de mentionner que, d'après l'expérience des auteurs, une fois qu'un examen potentiel et son processus associé ont été adoptés, en particulier s'ils proviennent d'une tierce partie, l'influence de l'organisme de réglementation pour obtenir des commentaires et/ou des modifications du contenu ou du processus de l'examen peut elle-même être limitée, voire inexistante. Dans le cas où des changements sont envisagés par la tierce partie, les coûts peuvent être importants, ce qui amène à se demander si ces ressources ne pourraient pas être mieux utilisées pour développer et maintenir le propre examen de l'ORD ou pour rejoindre une collaboration entre les organismes de réglementation, et ainsi créer un lien plus étroit entre l'examen et la pratique professionnelle telle qu'elle est dans la propre juridiction de l'ORD.

2.2 Adoption de l'examen potentiel avec modifications

Dans le cas où la similitude requise entre l'examen potentiel et l'examen de l'ORD n'est pas suffisante pour justifier une adoption complète, l'ORD peut chercher à adopter une partie du contenu et du processus de l'examen potentiel pour répondre à un besoin d'évaluation spécifique. Cette section décrit trois modifications possibles d'un examen potentiel pour répondre à des besoins spécifiques. Ces modifications se rapportent à des étapes particulières du processus d'élaboration de l'évaluation et décrivent donc ces étapes avec quelques détails techniques. Il convient de noter que, bien que ces changements soient possibles en principe, l'option de modifier le contenu ou le processus dépend de la volonté de l'entité responsable de l'examen potentiel.

Les trois options possibles examinées ici sont les suivantes :

- 1) Adopter des sous-ensembles du contenu potentiel de l'examen,
- 2) Modifier la contribution des éléments individuels à la note de passage ou à la note totale, et
- 3) Reclasser les éléments dans le CC de l'ORD.

Pour la première option, l'adoption de sous-ensembles du contenu de l'examen, il convient de noter que dans la section ci-dessus sur l'évaluation de la similitude entre l'examen potentiel et l'examen de l'ORD ou les éléments nécessaires pour en développer un (CC et autres), il a été avancé qu'il était plus probable que les différentes juridictions s'accordent sur la définition des aspects techniques d'une profession que sur celle des aspects non techniques. Par conséquent, il pourrait s'agir d'une stratégie rentable pour l'ORD d'adopter uniquement le contenu technique d'un examen potentiel et non la partie non technique. Si cette option était choisie, l'ORD devrait compléter l'examen potentiel par un contenu non technique développé soit en interne, soit par une tierce partie. Ce contenu pourrait être inclus dans le cadre d'une évaluation spécifique à une juridiction, telle qu'un examen sur le cadre juridique, le contexte et la déontologie de la profession.

L'un des problèmes posés par cette approche serait la complexité accrue de la notation de l'examen. Bien que l'on puisse s'attendre à ce que l'entité responsable de l'examen potentiel fournisse les notes des candidats et la note de passage pour la partie technique, l'ORD pourrait vouloir les combiner en une seule note avec les performances de la partie non technique. En outre, compte tenu de leur caractère distinct, l'ORD pourrait vouloir imposer des normes minimales pour les parties techniques et non techniques des examens. Ces processus ne sont pas particulièrement onéreux en soi, mais, intégrés dans différentes évaluations et différents systèmes de collecte de données, ils nécessiteraient des ressources supplémentaires pour une seule administration. De plus, si l'ORD décide que le contenu non technique d'un examen potentiel doit également être inclus, l'attribution de ce contenu à chaque élément du CC doit également être entreprise. Ce sujet sera abordé dans le cadre de la troisième option, ci-dessous.

La deuxième option consiste à modifier la relation entre la performance sur les items individuels et la décision de réussite ou d'échec afin qu'elle s'aligne mieux sur les normes de la juridiction de l'ORD. Cette option se justifie par le fait qu'elle reconnaît que des différences de normes peuvent exister entre la ou les juridictions utilisant l'examen potentiel et l'ORD, et qu'une meilleure harmonisation peut être obtenue soit en modifiant la note de passage de l'examen potentiel, soit en ne notant pas les éléments qui ne sont pas directement pertinents pour les normes de l'ORD.

La modification de la note de passage à la lumière des normes locales implique le processus de définition des normes, décrit plus en détail dans la section 4. À un niveau élevé, l'établissement de normes implique d'évaluer la probabilité qu'un candidat minimalement compétent réponde correctement au contenu de l'examen, un jugement qui implique un examen minutieux des normes d'accès à la pratique. Dans le cadre de la

stratégie proposée ici, des experts en la matière de l'ORD évalueraient le contenu potentiel de l'examen et porteraient des jugements sur sa difficulté en tenant compte des normes de pratique locales pertinentes afin de créer une note de passage spécifique à la juridiction. Bien que cette nouvelle note de passage soit basée sur le contenu créé par l'entité responsable de l'examen potentiel, les attentes concernant la compétence d'accès à la pratique seraient intégrées dans la nouvelle note de passage, ce qui lui permettrait de fonctionner plus efficacement dans la juridiction de l'ORD.

Une étape connexe consiste à éliminer des éléments spécifiques de la notation lorsque le contenu auquel ils se réfèrent n'est pas pertinent pour les normes de l'ORD. Un exemple spécifique concerne les processus d'assurance propres aux États-Unis qui pourraient faire partie des examens américains de physiothérapie ou d'orthophonie, présentés précédemment à titre d'illustration. Plutôt que de modifier l'examen lui-même, ce que l'entité responsable de l'examen potentiel pourrait être réticente à faire, la contribution de l'élément à la note totale serait mise à zéro et la note de passage serait réduite d'un montant approprié. Bien que les candidats soient toujours soumis aux questions, ils ne seraient évalués pour l'autorisation d'exercer que sur les questions qui sont pertinentes pour les normes de l'ORD. Il est à noter que l'inclusion d'items non notés est une pratique courante dans de nombreux programmes d'examens à enjeux élevés, car elle permet par ailleurs de recueillir des statistiques de performance sur les nouveaux items afin de déterminer leur qualité avant qu'ils n'affectent l'évaluation des candidats.

La troisième et dernière option de modification de l'examen décrite ici consiste à aligner plus explicitement le contenu de l'examen potentiel sur le CC de l'ORD. Cette étape consisterait à classer les questions de l'examen potentiel en fonction des compétences, des sous-compétences ou des indicateurs appropriés dans le CC, ce qui rendrait la pertinence de ces questions directement liée à la compétence de l'ORD. Cette étape présente plusieurs avantages. Tout d'abord, le fait de classer les questions de l'examen dans le cadre de l'ORD répondrait à la préoccupation selon laquelle les sujets de l'examen potentiel sont différents de celui de l'ORD ; les questions seraient explicitement liées au cadre de sa propre juridiction.

Deuxièmement, la classification des éléments selon le CC permettrait une évaluation rigoureuse de l'adhésion de l'examen potentiel au plan directeur présumé de l'ORD. Cette évaluation permettrait non seulement d'analyser le formulaire d'examen en question, mais aussi de préparer le terrain pour un document plus fondamental reliant les CC de l'examen potentiel et de l'ORD. Ce document serait d'une grande utilité lorsqu'il s'agirait de justifier le bien-fondé de l'examen potentiel dans l'ORD, car il rendrait explicite la manière dont les deux cadres s'articuleraient l'un par rapport à l'autre. En outre, l'identification d'écarts de contenu entre les deux cadres de référence, comme le montrent les questions alignées sur un seul des deux cadres, motiverait soit l'élaboration de questions supplémentaires, soit la nécessité d'évaluations supplémentaires, soit même le rejet de l'examen potentiel en tant que remplacement ou alternative à un examen construit en interne ou par le biais d'une collaboration entre des organismes de réglementation.

Comme le montrent clairement les étapes ci-dessus, qui pourraient être suivies pour adopter un examen potentiel avec des modifications, il n'y a pas d'option simple disponible. Chacune des trois étapes présentées ici, à savoir l'adoption de sous-ensembles de l'examen, la modification de la notation et/ou de la note de passage, et la reclassification des éléments dans les CC de l'ORD, nécessite des ressources internes importantes pour être mise en œuvre. On peut donc raisonnablement se demander si ces efforts en valent la peine par rapport à la création de son propre examen en interne ou par le biais d'une collaboration entre organismes de réglementation, en particulier lorsque même les options susmentionnées ne conduisent pas à un examen qui serait bien aligné sur le contexte de l'ORD. Dans la section suivante, une brève analyse sera fournie sur les besoins continus pour la maintenance de l'examen une fois que l'examen est adopté. Cette analyse montre que les besoins en ressources ne s'arrêtent pas à l'adoption de l'examen ; du temps et du personnel seront nécessaires pour garantir sa pertinence, sa défendabilité et son équité.

3. MAINTIEN ET AMÉLIORATION D'UN EXAMEN POTENTIEL APRÈS SON ADOPTION

Une fois qu'un examen potentiel a été adopté, en totalité ou en partie, l'ORD a l'obligation de vérifier qu'il continue à fonctionner efficacement dans le contexte d'une décision d'autorisation d'exercer. À moins que l'entité responsable de l'examen adopté ne soit engagée pour effectuer cette détermination au nom de l'ORD, cela nécessitera des ressources, de l'expertise et un effet de levier en termes de capacité. Il convient également de reconnaître que si l'entité responsable est chargée de cette fonction, elle se trouve dans une situation de conflit d'intérêts aiguë ; son approbation du processus existant lui permettra de poursuivre ses activités. C'est pourquoi l'ORD doit conserver un certain contrôle et une certaine surveillance lors de l'évaluation de la performance de l'examen adopté. Bien entendu, ce rôle nécessite des ressources.

Les responsabilités de l'ORD dans le maintien de la défendabilité après l'adoption relèvent de deux grandes catégories : la politique et le processus.

3.1 Politique

Sur le plan politique, la principale question concerne l'actualité du contenu des examens. D'une part, cette actualisation peut être liée à des normes de pratique, telles que des changements dans la technologie ou les normes de soins, en réponse à de nouvelles recherches ou technologies. Dans ce cas, ces changements peuvent s'appliquer de la même manière à l'entité responsable de l'examen adopté et à ses membres habituels, et le contenu de l'examen peut donc évoluer pour intégrer ces changements. D'autre part, l'actualisation pourrait être liée à des changements dans la législation ou à de nouvelles obligations concernant l'ORD. Dans ce cas, tout changement de contenu ne s'appliquerait qu'à l'ORD et des choix devraient être faits quant au contenu de l'examen nécessaire pour y répondre.

3.2 Processus

Du côté des processus, le maintien de la défendabilité du contenu de l'examen se rapporte à une ou plusieurs des étapes de développement de l'examen décrites plus en détail dans la Section 4. Les processus d'examen spécifiques susceptibles d'être impliqués dans la maintenance sont les suivants.

Premièrement, il se peut que l'entité responsable de l'examen adopté ou l'ORD apporte des modifications à leurs plans directeurs respectifs. Étant donné que, comme indiqué dans les sections 1 et 2, la similitude des programmes d'examen dépend en partie de l'alignement de ces plans, il peut être nécessaire de modifier les sujets et/ou la quantité de contenu de l'examen afin d'obtenir un alignement acceptable. Ces changements

peuvent impliquer l'élaboration d'un contenu supplémentaire, des modifications des règles de notation ou de la note de passage.

Si l'ORD n'a pas adopté l'ensemble de l'examen potentiel, il se serait engagé à apporter une partie du contenu à l'ensemble de l'examen. Dans ce cas, un nouveau contenu devra être créé en permanence afin de s'assurer qu'il reste d'actualité et qu'il n'est pas exposé lors d'administrations répétées. Une fois encore, ce processus est décrit plus en détail dans la section 4, mais une bonne règle empirique pour la taille de la banque d'items par rapport à la longueur du test est de 5:1, voire 10:1, en fonction de la fréquence d'administration. La création d'un nouveau contenu nécessite généralement la formation d'experts en la matière et la gestion du projet de développement d'éléments, y compris la création, la révision et les tests sur le terrain (voir Figure 1).

L'une des étapes les plus importantes pour maintenir la défendabilité d'un examen adopté est de vérifier la pertinence du contenu de l'examen par rapport à la juridiction et au contexte de l'ORD. Bien qu'un examen de la pertinence ait déjà été effectué (comme décrit dans les sections 1 et 2), une source clé d'information pour cette détermination peut être trouvée dans les réponses des candidats à ces questions. En analysant la difficulté et la discrimination à partir des données de réponse aux questions, il est possible d'obtenir des informations clés sur la correspondance entre les normes dans les juridictions qui utilisent l'examen. En particulier, si les questions s'avèrent beaucoup plus difficiles pour les candidats de l'ORD ou si la corrélation entre la performance sur certaines questions et le reste du test est faible, cela indique que le contenu des questions n'est pas bien compris par les candidats et donc que le contenu auquel elles font référence n'est peut-être pas pertinent pour eux.

Enfin, les notes de passage doivent être constamment ajustées pour compenser les différences de difficulté entre les différentes versions de l'examen. Lorsque l'ORD a adopté l'ensemble de l'examen, mais a choisi d'ajuster la note de passage pour qu'elle corresponde à sa juridiction, ou s'il a adopté l'examen potentiel en partie, des processus spécifiques doivent être mis en œuvre pour procéder à cet ajustement. Il peut s'agir d'établir des normes pour des questions particulières, d'utiliser la méthode de la mise en équivalence pour déterminer la norme de réussite pour de nouvelles questions, ou les deux à la fois (comme décrit dans la section 4). Dans tous les cas, ces processus nécessitent du temps et des ressources, mais ils sont essentiels au maintien de la défendabilité et de la validité des résultats d'examen. Cela dit, il est entendu que les organisations sont soumises à d'importantes contraintes en matière de ressources, tant financières qu'humaines. Par conséquent, il peut être tentant d'adopter ce qui peut sembler superficiellement être un examen équivalent pour sa propre juridiction. Ce choix doit être fait en tenant compte des implications et des conséquences.

4. ÉLABORATION D'UN EXAMEN D'ADMISSION EN INTERNE

La possibilité pour un organisme de réglementation (ou un groupe d'organismes de réglementation dans le cadre d'une collaboration) de développer son propre programme d'examen d'admission en interne est la voie privilégiée dans les cas où l'adoption d'un examen existant (en particulier d'une tierce partie ou d'une juridiction étrangère) n'est pas défendable ou réalisable. La plupart des professions réglementées au Canada exigent la réussite d'un examen pour l'obtention d'une autorisation d'exercer et accéder à la profession. Cela s'explique principalement par le fait qu'un organisme de réglementation ou un groupe d'organismes de réglementation peut s'assurer que le programme d'examen développé sous son contrôle répond aux besoins de ses parties prenantes au sein de sa juridiction, tant que le programme d'examen respecte les normes d'une évaluation psychométriquement défendable. En outre, l'élaboration d'un programme d'examen permet à l'organisme de réglementation de créer un programme financièrement autonome dans lequel les frais d'examen facturés aux candidats couvrent, dans bien des cas, le coût de l'élaboration et de la maintenance du programme.

Cette section présente une vue d'ensemble des étapes nécessaires à l'élaboration d'un programme d'examen de certification ou d'autorisation d'exercer à enjeux élevés en interne.

Il existe plusieurs normes internationalement reconnues dans le domaine de l'évaluation des enjeux, notamment :

- a) Normes de pratique du *testing* en psychologie et en éducation (Standards for Educational and Psychological Testing – SEPT)⁴
- b) *Directives de l'International Test Commission (ITC) pour la traduction et l'adaptation des tests*⁵
- c) *Normes d'accréditation des programmes de certification de la National Commission for Certifying Agencies (NCCA)*⁶
- d) ISO/IEC 17024 *Évaluation de la conformité – Exigences générales pour les organismes de certification de personnes*⁷

⁴ American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association. <https://www.apa.org/science/programs/testing/standards>
Version traduite en français : *Normes de pratique du testing en psychologie et en éducation*, traduit par l'OCCOPPQ sous la direction de Georges Sarrazin, Montréal, Institut de recherches psychologiques, 2003.

⁵ International Test Commission. (2017). *ITC Guidelines for Translating and Adapting Tests (Second Edition)* [www.InTestCom.org].
[Version traduite en français](#) élaborée pour la Société française de psychologie par Kamel GANA, Guillaume BROU, Nedjem Eddine Boudouda, Nicolas Calcagni, et Samia Ben Youssef.

⁶ En anglais : <https://www.credentialingexcellence.org/Accreditation/Earn-Accreditation/NCCA>

⁷ <https://www.iso.org/fr/standard/52993.html>

Les normes NCCA, les directives ITC pour la traduction et l'adaptation des tests et les normes ISO 17024 sont cohérentes avec les normes SEPT et ce document se concentrera donc sur les normes SEPT et leurs critères spécifiques.

Les normes d'évaluation fournissent un cadre pour la défendabilité psychométrique d'un programme d'examen et si un programme d'examen ne respecte pas les normes, il ne peut pas être mis en œuvre de manière défendable par une organisation. Par conséquent, si un programme d'examen potentiel ou interne ne répond pas aux normes de base applicables, l'organisation doit cesser d'envisager le programme d'examen potentiel ou revoir son examen interne.

Les normes SEPT sont généralement regroupées, pour des raisons de commodité, dans les principaux domaines du « cycle de vie » d'un examen à enjeux élevés, à savoir

1. Profil de compétences/Spécifications des tests
2. Spécifications de l'examen (Plan Directeur)
3. Interprétation du résultat
4. Données pour la validation
5. Variance non pertinente du construit et équité des tests
6. Analyse des items et des tests
7. Développement des items
8. Développement de l'examen
9. Administration de l'examen
10. Accommodements pour les tests
11. Établissement d'une note de passage
12. Notation et rapports

Pour l'élaboration d'un programme d'examen interne défendable sur le plan psychométrique, chacun des domaines susmentionnés doit respecter les normes de meilleures pratiques. Les domaines 1 et 2 de ces processus ont été décrits dans les sections précédentes du présent rapport. Les autres domaines sont décrits ci-dessous. Les normes SEPT relatives à chaque section figurent à l'[annexe A](#).

3. Interprétation du résultat

L'interprétation « directe » des résultats de l'examen, par exemple l'aptitude d'un candidat à assumer une pratique professionnelle indépendante dans sa profession, doit être correctement définie et documentée par le programme d'examen. Cela inclut la conception de l'examen, les processus d'élaboration, d'administration, d'analyse et surtout de communication des résultats. Dans le cas d'un examen basé sur des critères, où une « réussite » indique qu'un candidat peut être autorisé à exercer et qu'il est prêt à exercer sa profession de manière indépendante, cette interprétation des résultats de l'examen doit être étayée par des processus et une documentation défendables. De la même manière, si

un « échec » indique qu'un candidat n'est pas encore prêt à obtenir une autorisation d'exercer et qu'il n'est donc pas prêt à exercer sa profession de manière indépendante, cette interprétation des résultats de l'examen doit également être étayée.

4. Données pour la validation

Données pour la validation⁸ décrit les types de données qui devraient être recueillies pour aider à confirmer la validité du programme d'examen par une tierce partie. En conséquence, les normes pertinentes décrivent les différentes facettes de la collecte de données qui sont importantes pour la validation, y compris les caractéristiques des candidats composant l'échantillon, les possibilités de collecte de données et les types de mesures qui reflètent la validité à la fois au niveau de la tâche et du test.

5. Variance non pertinente de la construction et équité des tests

La variance non pertinente avec le construit (Construct Irrelevant Variance – CIV)⁹ est un terme psychométrique qui fait référence à la variation des résultats d'un test qui se produit indépendamment du trait ou de la caractéristique évaluée. Ces CIV ajoutent du « bruit » sous forme d'erreur à l'estimation de la capacité réelle d'un candidat. Parmi les exemples de CIV, on peut citer

- Questions d'examen mal construites¹⁰ qui provoquent la confusion des candidats et les empêchent de répondre aux questions de manière à démontrer leurs connaissances, compétences et aptitudes (Knowledge, Skills and Abilities – KSAs).
- Les candidats utilisent des stratégies de simulation pour « trouver » les réponses à des questions mal construites.
- Une personne très compétente peut avoir de mauvais résultats à un test parce qu'elle est tombée malade. Dans ce cas, la maladie serait considérée comme une source de CIV.
- Un bruit de construction distrayant (irrégularité de test) était présent dans un centre de test pendant que les candidats passaient un examen. Dans ce cas, le bruit de construction distrayant pouvait gêner les candidats dans la démonstration de leurs « véritables » connaissances, compétences et aptitudes (Knowledge, Skills and Abilities – KSAs) lors de l'examen et pouvait donc être une source de CIV.

⁸ En anglais : Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Medical Education. 2004; 38:327-333.

⁹ En anglais : Haladyna T. M., Downing S. M. (2004). Construct-irrelevant variance in high-stakes testing". Educational Measurement: Issues and Practice, 23(1), 17-27. <https://doi.org/10.1111/j.1745-3992.2004.tb00149.x>

¹⁰ En anglais : Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ. 2005; 10:133-143.

- Si un examen est proposé en anglais et en français, mais que la traduction française de l'examen n'est pas suffisamment précise. Dans ce cas, les candidats français risquent de ne pas être aussi performants qu'ils l'auraient été si l'examen avait été correctement traduit en français, et la traduction de l'examen est donc une source de CIV. L'inverse est également vrai.

6. Analyse des items et des tests

L'analyse des items (questions) et du test est l'étape au cours de laquelle les analyses statistiques au niveau des questions et de l'examen sont effectuées pour s'assurer que les questions et l'examen dans son ensemble répondent aux normes de défendabilité.

Une fois les items développés et examinés, ils doivent être administrés à des échantillons représentatifs afin d'obtenir des statistiques sur les items et les tests permettant d'évaluer la performance statistique des items expérimentaux (pas encore opérationnels). La meilleure pratique pour administrer des items expérimentaux est d'utiliser un processus appelé test sur le terrain. Plusieurs méthodes de test sur le terrain sont couramment utilisées dans le secteur des évaluations à enjeux élevés, la plus courante étant d'inclure un petit nombre (par exemple, 10 à 15 % du nombre d'items opérationnels notés) d'items expérimentaux non notés dans les formulaires d'examen que les candidats passent. Comme les candidats ne savent pas quelles sont les questions expérimentales, ils répondent aux questions expérimentales de la même manière qu'aux questions opérationnelles notées de l'examen. De cette manière, des données précises et représentatives sont collectées sur les questions expérimentales, ce qui permet d'analyser les questions et le test. Il est important de noter que si l'on ne recueille pas d'informations sur l'analyse des items et des tests, il n'est pas possible de déterminer avec précision la performance des items et des examens qui seront composés de ces items. En fait, si aucune analyse des questions et des tests n'est effectuée pour un programme d'examen, il n'y a vraiment aucun moyen de s'assurer que les questions et l'examen dans son ensemble fonctionnent statistiquement comme ils le devraient (c'est-à-dire que l'examen du contenu en l'absence de données statistiques n'est pas suffisant pour évaluer la performance acceptable des questions).

Il existe deux approches communes pour tester les éléments d'un examen : la théorie classique des tests (TCT) et la théorie de la réponse aux items (TRI).

Théorie classique des tests (TCT)¹¹ est l'approche la plus couramment utilisée pour l'analyse statistique des questions de test, principalement en raison de la simplicité du calcul et de l'interprétation des statistiques, ainsi que du fait que la TCT nécessite généralement des échantillons de plus petite taille que la TRI. Les analyses d'items de la TCT comprennent généralement les statistiques suivantes :

¹¹ En anglais : Crocker, L. et Algina, J. (1986) Introduction to Classical and Modern Test Theory. Harcourt, New York.

- *Valeur p* : La proportion de candidats ayant choisi la bonne réponse à la question (par exemple, une valeur p de 0,65 indique que 65 % des candidats ayant répondu à la question ont choisi la bonne réponse).
- *Corrélation item-total (discrimination)* : Corrélation entre la note de l'item et la note totale de l'examen. Il s'agit généralement d'une corrélation point-bisérial qui est soit corrigée (la note de l'item est retirée du calcul de la note totale), soit non corrigée. Des corrélations point-bisérial corrigées de 0,2 ou plus indiquent généralement que la question permet une bonne discrimination entre les candidats ayant obtenu des scores élevés à l'examen, des scores modérés à l'examen et des scores faibles à l'examen, car les candidats ayant obtenu des scores élevés à l'examen ont tendance à répondre correctement à la question plus souvent que les candidats ayant obtenu des scores modérés ou faibles. Des corrélations item-total négatives pour un item peuvent indiquer que la question a été mal formulée.
- *Ventilation des options de réponse* : Selon le type de question, une répartition de la proportion de candidats au total, ventilée par score global à l'examen (« élevé », « moyen », « faible »), est calculée. Ces informations appuient la valeur p et les statistiques de discrimination, indiquant si une question peut dérouter les candidats (entraînant potentiellement une valeur de discrimination plus faible) ou si elle donne les résultats escomptés.
- *Le temps consacré à chaque question* : Pour les examens informatisés, la plateforme d'examen recueille des informations sur le temps que les candidats consacrent à chaque question, en minutes ou en secondes. Bien qu'il ne s'agisse pas techniquement d'une statistique du TCT, cette information est utile pour évaluer la performance des questions, car les questions auxquelles les candidats consacrent beaucoup de temps ou presque pas du tout doivent être examinées de près.

La Figure 1 présente un exemple d'analyse d'un item du TCT pour une question qui répond à tous les critères statistiques d'un programme d'examen. La valeur p n'est ni trop élevée ni trop basse, puisqu'elle est de 0,68 (68 % des candidats ont choisi la bonne réponse, « C », pour la question), la statistique de discrimination est supérieure à 0,2, ce qui indique une corrélation positive acceptable entre les scores de la question et les scores de l'examen, et le temps moyen total consacré à la question est de 1,2 minute. L'analyse de la question montre également où les candidats sont allés lorsqu'ils ont sélectionné les distracteurs (alternatives incorrectes « A », « B », et « D' ») avec approximativement la même proportion de candidats sélectionnant chacun des distracteurs. La partie surlignée en jaune montre également que les candidats ayant obtenu des résultats élevés à l'examen (« 1. Élevés » : 79,1 %) sélectionnent la réponse correcte « C » dans des proportions plus élevées que les candidats ayant obtenu des résultats moyens à l'examen (« 2. Moyens » : 68,7 %) et les candidats ayant obtenu des résultats faibles à l'examen (« 3. Faibles » : 51,2 %).

Figure 1 : Exemple d'analyse d'un item de la théorie classique des tests (TCT) pour un item qui répond aux normes statistiques

Meets standards?	Item ID	Key	N	p-value (difficulty)	Item-total correlation (discrimination)	Average total time spent on item (minutes)
YES	24133	C	231	0.68	0.233	1.2
	Ability	A	B	C	D	No response
	High	6.5%	9.0%	79.1%	5.5%	-
	Mid	11.1%	10.1%	68.7%	9.4%	0.6%
	Low	15.6%	13.7%	51.2%	16.6%	2.9%
	TOTAL	11.1%	10.5%	67.5%	10.0%	0.9%

La Figure 2 présente un exemple d'analyse d'un item du TCT pour une question qui ne répond pas à tous les critères statistiques d'un programme d'examen. La valeur p est très faible et s'élève à 0,22 (seulement 22 % des candidats ont choisi la bonne réponse de « C » pour la question), la statistique de discrimination est inférieure à 0,200, à 0,023, ce qui indique une corrélation positive inacceptable entre les scores de la question et les scores de l'examen, et le temps moyen total passé sur la question de 3,2 minutes est beaucoup plus long que le temps moyen passé par question pour ce programme d'examen qui est d'une minute. L'analyse de l'item montre également où les candidats sont allés lorsqu'ils ont sélectionné les distracteurs (alternatives incorrectes « A », « B », et « D ») avec des candidats sélectionnant les options « A » et « D » dans des proportions plus élevées par rapport à la bonne réponse. Cela peut indiquer que les candidats ne connaissaient pas la matière testée par la question alors qu'ils devraient la connaître, ou que la question comporte des lacunes qui incitent les candidats à sélectionner les distracteurs au lieu de la bonne réponse, ou encore que la question est tout simplement difficile pour les candidats. Une évaluation du contenu de l'item en conjonction avec les statistiques d'analyse de l'item peut aider à résoudre ce qui se passe avec l'item. Il est important de noter que les questions ne doivent pas nécessairement être rejetées sur la base des seules statistiques de la question ; un examen du contenu par un expert en la matière, avec les informations statistiques à l'appui, permettra de comprendre pourquoi la question a donné ces résultats. La zone surlignée en jaune montre également que les candidats ayant obtenu un score élevé à l'examen (« 1. Élevé » : 33,1 %) sélectionnent la bonne réponse « C » dans des proportions plus élevées que les candidats ayant obtenu un score moyen à l'examen (« 2. Moyen » : 20,6 %) et les candidats ayant obtenu un score faible à l'examen (« 3. Faible » : 17,6 %), ce qui explique pourquoi la corrélation item-total n'est pas plus faible (la bonne réponse attire une plus grande proportion de candidats ayant obtenu un score élevé à l'examen que de candidats ayant obtenu un score faible à l'examen).

Figure 2 : Exemple d'analyse d'un item de la théorie classique des tests (TCT) pour un item qui ne répond pas aux normes statistiques

Meets standards?	Item ID	Key	N	p-value (difficulty)	Item-total correlation (discrimination)	Average total time spent on item (minutes)
NO	33421	C	231	0.22	0.023	3.2
	Ability	A	B	C	D	No response
	High	35.4%	8.7%	33.1%	22.0%	0.8%
	Mid	36.3%	10.8%	20.6%	30.4%	1.8%
	Low	27.2%	25.0%	17.6%	18.4%	11.8%
	TOTAL	34.5%	13.0%	22.2%	26.8%	3.5%

Les analyses de tests basées sur la théorie classique des tests ont tendance à consister en des statistiques descriptives pour l'examen, y compris les mesures de tendance centrale (par exemple, moyenne, médiane), de variation (par exemple, écart-type) et des informations sur les fourchettes (par exemple, les scores maximum et minimum obtenus). La fiabilité de la cohérence interne est généralement la formule de Kuder-Richardson 20 – KR-20 – (pour les questions à notation dichotomique) ou l'alpha de Cronbach (pour les questions à notation polytomique) et varie de zéro à un, les valeurs plus proches de un indiquant une plus grande fiabilité de la cohérence interne. La Figure 3 présente un exemple de rapport d'analyse de test.

Figure 3 : Exemple de rapport d'analyse d'un test de la théorie classique des tests (TCT)

Test analysis report

Exam name: Example Examination June 2023	Exam owner: Susan Smith
Report generation date/time: June 30, 2023, 0800	Report filters: Date range June 7-29, 2023
Number of candidates: 231	Number of items: 100
Mean: 65.2%	Standard deviation: 10.3%
Median: 67.3%	Max score: 100%
Mode: 62%	Min score: 36%
Standard error of measurement: 8.22	Internal consistency reliability: 0.700

La théorie de la réponse à l'item (TRI)¹² est une autre approche utilisée pour analyser statistiquement les questions d'examen. La TRI adapte la performance des candidats à un modèle de réponse à l'item qui utilise entre un et trois paramètres pour définir la performance des items.

Pour un modèle de réponse à l'item à un paramètre, seul un paramètre b est utilisé. Il s'agit de la difficulté de la question sur une échelle normale standard (appelée Thêta – le continuum de capacité pour le domaine de contenu évalué) allant approximativement de -3,5 à +3,5. Une question de difficulté moyenne aura une valeur de paramètre b proche

¹² En anglais : Hambleton, R. K., Swaminathan, H. (1985). Item response theory: Principles and applications. Boston : Kluwer-Nijhoff.

de zéro, tandis qu'une question extrêmement difficile aura une valeur de paramètre b positive élevée (par exemple, 3,1) et des questions extrêmement faciles auront une valeur de paramètre b négative faible (par exemple, -3,4).

Un modèle de réponse à l'item à deux paramètres ajouterait une valeur de paramètre a (allant de 0 à 2) qui définit la discrimination de l'item, plus le nombre est élevé, plus la discrimination de l'item est grande.

Enfin, un modèle à trois paramètres ajoute un paramètre c aux paramètres b et a . Le paramètre c est parfois appelé « paramètre de pseudo-évaluation ». Le paramètre c fournit une mesure de la probabilité que les candidats peu doués choisissent la bonne réponse à la question, comprise entre 0 et 1 (bien qu'elle se situe de manière réaliste entre 0 et 0,4). La performance de l'item dans la TRI est modélisée visuellement à l'aide de courbes caractéristiques de l'item (CCI) qui ressemblent à une courbe en S où le paramètre b est la position de l'item sur l'axe des x (difficulté/aptitude appelée θ), le paramètre a est la pente de la CCI et le paramètre c est l'intersection de la CCI et de l'axe des y .

Les analyses de test basées sur la TRI ont tendance à se concentrer sur les fonctions d'information du test, qui indiquent où l'on trouve le plus et le moins d'informations sur le test (et donc où l'on trouve le moins d'erreurs de mesure sur le continuum aptitude/performance). Les concepteurs de tests peuvent déterminer la forme de la fiabilité de l'examen en sélectionnant les éléments qui fournissent le maximum d'informations (précision) en différents points du continuum d'aptitudes (axe des x , θ) dans les domaines où l'on a besoin du maximum d'informations pour différencier les candidats compétents de ceux qui ne le sont pas encore (par exemple, au niveau d'une note de passage - *cut score*).

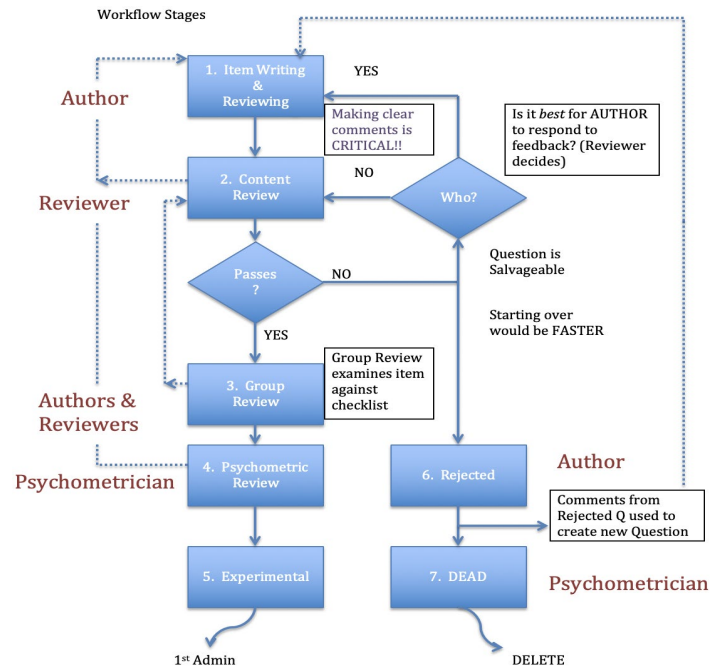
7. Développement des items

Le développement d'items ou de questions est le processus de création de questions d'examen et d'alimentation de la banque d'items requis pour le programme d'examen. Pour que les processus de développement des questions répondent aux normes et soient donc considérés comme psychométriquement défendables, il est nécessaire qu'un programme d'examen aligne le développement des questions sur le plan directeur d'examen et suive les meilleures pratiques en matière de développement des questions, et qu'il y ait un processus rigoureux d'examen et d'approbation.

Si la traduction des éléments est nécessaire, le processus de traduction suit les meilleures pratiques et la traduction de chaque élément de la banque est vérifiée par un traducteur professionnel.

La Figure 4 présente un exemple de processus de développement d'items qui répondrait aux normes s'il était mis en œuvre dans le cadre d'un programme typique d'examen à enjeux élevés avec notation objective.

Figure 4 : Exemple de processus d'élaboration d'un item



Des pratiques saines de développement et d'examen des questions contribueront à atténuer le risque de sous-représentation des concepts¹³, qui se produit lorsque le contenu de l'examen ne reflète pas les connaissances, compétences et aptitudes (KSAs) essentielles pertinentes, ce qui dégrade la validité du programme d'examen. Voici quelques exemples de sous-représentation des concepts :

- Contenu considéré comme trivial ou futile en ce sens qu'il n'est pas important pour la mesure de la construction.
- Contenu qui dépend de la mémorisation par cœur des informations par le candidat.
- Le nombre de questions d'un examen n'est pas suffisant pour obtenir une précision de mesure suffisante (qui peut être estimée par des mesures de fiabilité de la cohérence interne telles que le KR-20/l'alpha de Cronbach).
- Représentation incorrecte d'item provenant de zones de plans directeurs ou autres violations des spécifications de l'examen.
- Les programmes d'éducation qui « enseignent pour le test » préparent les candidats à passer un examen plutôt que de se concentrer sur l'amélioration des compétences des candidats dans le domaine testé.

¹³ En anglais : Chen, Y., Li, I. et Chason, W. (2018). Construct underrepresentation. In The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation (Vol. 4, pp. -). SAGE Publications, Inc. <https://doi.org/10.4135/9781506326139>

8. Développement de l'examen

Le développement de l'examen¹⁴ est le processus d'assemblage des questions de haute qualité qui ont été soumises à tous les processus de révision des questions et ont été acceptées pour une utilisation opérationnelle sur un formulaire d'examen. Les types de questions vont des questions à choix multiples (QCM) et à sélection multiple (MSQ), notées objectivement, aux tâches notées par jugement, telles que les réponses écrites et les activités basées sur la performance. Le choix du type de question dépend de la compétence à évaluer ; lorsque la compétence est axée sur les aptitudes cognitives, les QCM et/ou les QMS sont souvent appropriés. Lorsque la compétence exige des performances, notamment en matière de communication, de démonstration ou de création, une tâche plus complexe et probablement notée avec discernement sera nécessaire.

Les spécifications de l'examen sont utilisées pour créer les formulaires d'examen, y compris les exigences relatives au plan directeur, au type d'item, aux statistiques de difficulté de l'item, aux statistiques de discrimination de l'item, au temps passé sur la question, aux ennemis de l'item et aux exigences relatives au contenu de la question. Voici un exemple de contraintes qui peuvent être requises pour un programme d'examen :

- Nombre de formulaires parallèles : Deux formulaires.
- Ampleur de l'examen : 150 items.
- Durée de l'examen : 3 heures
- Difficulté de l'item : Les items doivent avoir des valeurs p comprises entre 0,35 et 0,90 avec une moyenne sur l'ensemble du formulaire proche de 65 %.
- Discrimination des items : Les items doivent avoir des indices bisérialux de points corrigés supérieurs à 0,20.
- Couverture du plan directeur : Nombre de questions par catégorie du plan directeur.
- Couverture du niveau cognitif : Nombre de questions qui évaluent les « connaissances » par rapport aux « compétences ».
- Images des objets : Les articles comportant des images doivent être équilibrés dans la catégorie des plans directeurs.
- Passages : Nombre de passages équilibrés dans l'ensemble de l'examen.
- Nombre de questions associées aux passages : 3 à 6 questions par passage.
- Amis d'items : Les éléments qui doivent être présentés sur un formulaire d'examen sont liés les uns aux autres.
- Ennemis d'items : Certains objets ne peuvent pas être présentés l'un avec l'autre sur un formulaire d'examen.

¹⁴ En anglais : Downing, S. M., & Haladyna, T. M. (Eds.). (2006). Handbook of test development. Lawrence Erlbaum Associates Publishers.

9. Administration de l'examen

L'administration de l'examen est le processus par lequel les candidats passent l'examen dans des conditions où ils ont la possibilité de démontrer leurs connaissances, leurs compétences et leurs aptitudes sur le site avec le moins d'influences externes possible. Pour garantir la validité du programme d'examen, il est essentiel de s'assurer que le contenu de l'examen est sécurisé et que le matériel, les notes et les résultats sont traités en toute sécurité. La sécurisation de tous les aspects de l'examen permet de s'assurer que les candidats démontrent réellement leurs compétences lors de l'examen et que la banque d'examens n'est pas exposée aux candidats en attente de passer l'examen à l'avenir. En outre, l'administration défendable de l'examen nécessite la sécurisation du contenu des questions afin de limiter l'exposition des questions ainsi que la sécurisation de l'examen pendant l'administration afin de s'assurer que les candidats ne copient pas les réponses les uns des autres (c'est-à-dire qu'il n'y a pas de collusion). La validité du programme d'examen serait compromise si le contenu des questions était exposé (car les candidats pourraient utiliser ces questions exposées pour gonfler artificiellement leur performance à l'examen en ayant été préalablement exposés au contenu de l'examen) ou si les candidats s'entendaient pour augmenter artificiellement leur performance en copiant les réponses d'autres candidats.

10. Accommodements de tests

Un accommodement de test est une modification de l'administration standard d'un examen pour tenir compte d'une limitation fonctionnelle liée aux droits de la personne¹⁵. Les candidats peuvent bénéficier d'accommodements de test en fonction de la nature de leurs limitations fonctionnelles et des besoins d'aménagements associés. Un programme d'examen doit documenter et fournir des accommodements de test dans le cadre des exigences d'un programme d'examen défendable.

Parmi les exemples d'accommodements que les candidats peuvent généralement demander, on peut citer la prolongation de la durée de l'examen, les pauses, l'accès à des équipements d'assistance, à du personnel d'assistance ou à des articles personnels (par exemple, des tire-laits pour l'allaitement).

Si les accommodements de test pour un candidat handicapé ne sont pas correctement mis en place, le candidat qui a besoin de ces accommodements peut ne pas être en mesure de démontrer ses « véritables » connaissances, compétences et aptitudes (KSAs). Dans ce cas, l'absence d'accommodement introduit la CIV, comme indiqué dans la section précédente « V. Variance non pertinente pour la construction et équité des tests ».

¹⁵ <https://mcc.ca/fr/examens-et-evaluations/demandes-relatives-aux-examens/accommodements-dexamen/>

11. Établissement d'une note de passage

La note de passage de l'examen est l'élément qui formalise la classification des candidats entre ceux qui réussissent et ceux qui échouent, ce qui est l'objectif même de l'examen d'admission. Un programme d'examen peut faire tout ce qu'il faut lors de la conception du programme d'évaluation, mais si le programme d'examen ne parvient pas à identifier le « bon » niveau de performance qui définit la compétence d'entrée dans la pratique, la légitimité des décisions prises est compromise. Les programmes d'examen qui fixent des seuils de performance de manière arbitraire (par exemple, 50 %) et qui, par conséquent, ne respectent pas des bonnes pratiques, ne peuvent être considérés comme défendables.

Les approches en matière d'établissement de normes¹⁶ diffèrent selon les types d'examens concernés. Par exemple, les méthodes diffèrent selon qu'il s'agit d'examens à choix multiples ou d'examens basés sur la performance, tels que les ECOS (examens cliniques objectifs structurés). Il existe deux types principaux de méthodes de normalisation : les méthodes centrées sur le test et les méthodes centrées sur le candidat.

- La normalisation centrée sur le test se concentre sur le contenu du test lui-même, tel que l'item ou la tâche, afin de déterminer les normes minimales. Une approche typique centrée sur le test impliquerait une évaluation par des experts en la matière de la difficulté probable des questions et des tâches pour un candidat qui atteint tout juste la norme minimale. Cette évaluation ne porterait que sur la question ou la tâche elle-même et sur une analyse de la difficulté à la compléter ou à la résoudre. Voici quelques exemples d'approches de normalisation centrées sur les tests :
 - Méthode d'Angoff modifiée
 - Méthode de mise en signet (Bookmarking)
 - Méthode Nedelsky
- Les méthodes centrées sur le candidat exigent que les examinateurs comparent la performance du candidat aux normes minimales, puis identifient la note de test qui différencie le mieux les performances jugées réussies et non réussies. Par exemple, les examinateurs peuvent fournir des notes sur des facettes individuelles de la performance et classer la performance de manière globale comme satisfaisant ou non aux normes minimales. Le total des notes individuelles qui représente le mieux le seuil entre les performances jugées réussies et non réussies est adopté comme note de passage pour la tâche. Voici quelques exemples d'approches de fixation de normes centrées sur l'examineur :
 - Méthode de régression des limites
 - Méthode des groupes contrastés
 - Méthode Hofstee

¹⁶ En anglais : Thum, Yeow Meng. (2003). Setting Performance Standards: Concepts, Methods, and Perspectives. Applied Psychological Measurement - APPL PSYCHOL MEAS. 27. 305-307.

12. Notation et rapports

La notation¹⁷ et l'établissement de rapports sont le processus par lequel les psychométriciens prennent les notes et calculent les notes des candidats sur la base d'une approche de notation solide. Des approches systématiques et des logiciels (Excel, SPSS, R, etc.) sont généralement utilisés pour noter tous les candidats afin de garantir l'exactitude et la reproductibilité des notes et de la production des rapports. Les rapports des candidats sont conçus selon les normes des meilleures pratiques afin de fournir une transparence et un retour d'information appropriés (par exemple, des rapports de diagnostic pour les candidats qui échouent). Les rapports destinés aux parties prenantes concernant la performance des correcteurs, des stations et de l'examen dans son ensemble sont également généralement produits afin que l'organisation (entité responsable de l'examen, organismes de réglementation, entités de surveillance et autres) puisse utiliser les informations de chaque session pour surveiller et, le cas échéant, améliorer le programme d'examen.

L'objectif d'un examen est de découvrir quelque chose de significatif sur la possession d'un ou de plusieurs attributs (tels que la compétence d'accès à la pratique) par un individu ou un groupe. Pour ce faire, l'examen comprend des tâches qui représentent des preuves de la possession de ces attributs. Ces preuves sont ensuite combinées en scores qui représentent le degré de possession de l'attribut ou des attributs. Le processus consistant à traduire les réponses d'un candidat en preuves de la possession de l'attribut ou des attributs est communément appelé notation, tandis que la communication du degré de possession de l'attribut ou des attributs est appelée rapport. Les tâches sont généralement classées en deux catégories : les tâches notées objectivement et les tâches notées par jugement :

- Les tâches notées objectivement peuvent être évaluées sans expertise en la matière. L'exemple le plus courant de ce type de tâche est la question à choix multiples, qui demande aux candidats de sélectionner une option parmi un petit nombre d'options prédéterminées comme réponse à une question. La notation de ces questions ne nécessite qu'une comparaison entre l'option sélectionnée par le candidat et celle désignée comme correcte (par exemple, le candidat sélectionne « A » et la réponse correcte à l'item est « A » ; le candidat reçoit donc un point pour avoir sélectionné la bonne réponse à la question). Lorsque le candidat sélectionne l'option correcte, la question est notée comme correcte (par exemple, score de 1), et incorrecte dans le cas contraire (par exemple, score de zéro). Il est important de noter que ce processus de comparaison peut être effectué par une personne ne possédant aucune expertise dans le domaine, ou par une machine (c'est pourquoi les tâches notées objectivement sont parfois appelées « notées par ordinateur »). Parmi les autres exemples de tâches à notation objective, on peut citer la sélection multiple, dans laquelle plusieurs

¹⁷ En anglais : Thissen, D., & Wainer, H. (Eds.). (2001). Test scoring. Lawrence Erlbaum Associates Publishers.

alternatives peuvent être correctes pour une seule question, le remplissage de cases blanches, dans lequel un petit nombre de mots ou de chiffres sont comparés à ceux désignés comme corrects, et les tâches basées sur l'image ou les tâches dites « hotspot », qui demandent aux candidats de cliquer sur la zone pertinente d'un diagramme en fonction de l'invitation lancée pour la question. Chacun de ces types de questions ne nécessite aucune expertise en la matière pour être noté correctement une fois que la réponse correcte à la question a été identifiée ou déterminée.

- Les tâches notées par jugement requièrent des connaissances dans le domaine concerné afin de noter les performances avec précision. Ces tâches requièrent généralement des réponses plus ouvertes que celles qui sont notées objectivement. Par conséquent, les candidats disposent d'une grande flexibilité pour répondre à ces tâches, ce qui se traduit par un plus large éventail de réponses potentielles. Un exemple marquant de ce type de tâche est la question à développement, dans laquelle un ou plusieurs paragraphes sont requis pour répondre à une question ou à un problème présenté dans une question-guide. Toutefois, la réponse ne doit pas nécessairement être textuelle ; les dessins et les diagrammes, les enregistrements audio ou les performances en temps réel telles que celles requises pour les ECOS sont autant de réponses autorisées pour certaines tâches notées par jugement. Ces tâches sont notées par des experts en la matière qui comparent la performance ou le produit à des critères de qualité préétablis, en attribuant une note qui correspond le mieux à la description de la performance à ce niveau.

Un autre aspect de la notation consiste à garantir la comparabilité des notes entre les différentes administrations de l'examen. La cohérence lors de l'élaboration des formulaires d'examen est essentielle pour garantir le parallélisme de plusieurs formulaires d'examen. Pour ce faire, il convient de respecter les exigences des spécifications de l'examen pour chaque forme d'examen. La méthode d'équilibrage statistique¹⁸ est également généralement utilisée pour garantir la comparabilité des résultats d'une session à l'autre. Un modèle courant consiste à utiliser un ensemble d'items d'ancrage qui sont communs entre les versions et les sessions d'administration de l'examen. Ces items d'ancrage servent d'estimation de la capacité de la cohorte de candidats à passer l'examen qui, en conjonction avec la performance des items uniques de chaque version d'examen, permet au logiciel statistique de déterminer si des ajustements aux scores sont nécessaires pour s'assurer que les candidats qui passent une version légèrement plus difficile de l'examen ne sont pas désavantagés.

¹⁸ En anglais : Boucher, Thomas. (2007). Test Equating, Scaling, and Linking: Methods and Practices par Michael J. Kolen; Robert L. Brennan. Journal of the American Statistical Association.

Le rapport¹⁹ est le processus de communication des résultats aux parties prenantes, ce qui inclut non seulement le candidat (par exemple, « Réussite » ou « Échec »), mais aussi les parties prenantes régionales (par exemple, les résultats par zone géographique), les établissements d'enseignement (par exemple, les taux de réussite par programme postsecondaire), les organismes de réglementation, les entités de surveillance et d'autres encore. Par conséquent, les rapports doivent être facilement accessibles et présenter aussi clairement que possible ce que l'examen a révélé sur la possession de l'attribut ou des attributs par des individus ou des groupes, tout en évitant les possibilités d'interprétation erronée.

Les rapports sont généralement classés en deux catégories : les rapports fondés sur des normes et les rapports fondés sur des critères. Les rapports fondés sur une norme établissent des comparaisons avec les performances d'un groupe de référence, par exemple, le rapport sur le percentile. Dans cet exemple, la signification (et la conséquence) d'un résultat particulier découle de sa comparaison avec les performances des autres (par exemple, « Le candidat se situe dans le 95e percentile à l'examen de mathématiques » indique que le candidat a obtenu de meilleurs résultats que 95 % de tous les candidats qui ont passé l'examen de mathématiques). Dans le contexte de l'admission à une profession, la comparaison entre les performances des candidats diplômés d'un établissement d'enseignement donné et celles d'un échantillon national constitue un exemple de rapports référencés par rapport à des normes. En revanche, les rapports fondés sur des critères comparent les performances d'un individu ou d'un groupe à une norme, ou à un niveau de performance représentant la possession probable de certaines caractéristiques. Un exemple marquant d'une telle norme est la réussite/échec, et par conséquent, tout test conçu pour prendre une décision de réussite/échec serait considéré comme fondé sur des critères. Dans la pratique, les rapports comportent souvent des aspects liés à la fois aux normes et aux critères. Par exemple, un test fondé sur des critères (évaluation critériée), tel qu'un examen d'admission, peut également inclure les taux de réussite par administration (fondés sur la norme).

Les informations à inclure ou à ne pas inclure dans les rapports varient grandement d'un programme d'examen à l'autre et nécessitent généralement un examen des rapports par un psychométricien expérimenté afin de déterminer si les rapports sont défendables.

¹⁹ En anglais : Frey, B. (Ed.) (2018). The SAGE encyclopedia of educational research, measurement, and evaluation. (Vol. 1-4). SAGE Publications, Inc., <https://doi.org/10.4135/9781506326139>

ANNEXE A - CERTAINES NORMES SEPT APPLICABLES

Certaines des normes SEPT applicables à chaque domaine de l'élaboration d'un programme d'examen de certification ou d'autorisation d'exercer à enjeux élevés.

1. Profil de compétences/spécifications du test et 2. Spécifications de l'examen (plan directeur)

- SEPT 1.1 : Le concepteur du test doit indiquer clairement comment les résultats du test doivent être interprétés et donc utilisés. La (les) population(s) à laquelle (auxquelles) le test est destiné doit (doivent) être clairement délimitée(s), et le(s) construit(s) que le test est censé évaluer doit (doivent) être clairement décrit(s).
- SEPT 4.1 : Les spécifications du test doivent décrire le(s) objectif(s) du test, la définition du construit ou du domaine mesuré, la population de candidats prévue et les interprétations pour les utilisations prévues. Les spécifications doivent inclure une justification des interprétations et des utilisations des résultats du test pour le(s) objectif(s) prévu(s).
- SEPT 4.2 : Outre la description des utilisations prévues du test, les spécifications du test doivent définir le contenu du test, la durée proposée, les formats des items, les propriétés psychométriques souhaitées des items et du test, ainsi que l'ordre des items et des sections. Les spécifications du test doivent également préciser le temps alloué au test, les instructions pour les participants, les procédures à utiliser pour l'administration du test, y compris les variations autorisées, tout matériel à utiliser, ainsi que les procédures de notation et de rapport. Les spécifications des tests informatisés doivent inclure une description des exigences en matière de matériel et de logiciel.
- SEPT 4.12 : Les développeurs de tests doivent documenter la mesure dans laquelle le domaine de contenu d'un test représente le domaine défini dans les spécifications du test.
- SEPT 5.20 : Si les spécifications d'un test sont modifiées d'une version à l'autre, ces changements doivent être identifiés et une indication doit être donnée que les scores convertis pour les deux versions peuvent ne pas être strictement équivalents, même si des procédures statistiques ont été utilisées pour relier les scores des différentes versions. Lorsque des changements substantiels interviennent dans les spécifications du test, les scores doivent être rapportés sur une nouvelle échelle, ou une mention claire doit être fournie pour avertir les utilisateurs que les scores ne sont pas directement comparables à ceux des versions antérieures du test.
- SEPT 4.8 : Le processus de révision des tests doit inclure des analyses empiriques et/ou le recours à des juges experts pour réviser les items et les critères de notation. Lorsque des juges experts sont utilisés, leurs qualifications, leurs expériences pertinentes et leurs caractéristiques démographiques doivent être documentées, ainsi que les instructions et la formation au processus d'examen des items que les juges reçoivent.

3. Interprétation des notes

- SEPT 1.2 : Une justification doit être présentée pour chaque interprétation prévue des résultats de tests pour une utilisation donnée, ainsi qu'un résumé des preuves et de la théorie concernant l'interprétation prévue.
- SEPT 1.3 : Si la validité d'une interprétation courante ou probable pour un usage donné n'a pas été évaluée, ou si une telle interprétation n'est pas cohérente avec les données disponibles, ce fait doit être clairement indiqué et les utilisateurs potentiels doivent être fortement mis en garde contre les interprétations non étayées.
- SEPT 1.5 : Lorsqu'il est clairement indiqué ou sous-entendu que l'interprétation d'un score de test recommandé pour une utilisation donnée aboutira à un résultat spécifique, la base de l'attente de ce résultat doit être présentée avec des preuves pertinentes.
- SEPT 1.7 : Si la performance d'un test ou une décision prise à partir de celle-ci est censée ne pas être affectée par la pratique et l'entraînement, alors la propension de la performance du test à changer avec ces formes d'enseignement doit être documentée.
- SEPT 4.15 : Les instructions relatives à l'administration du test doivent être présentées avec suffisamment de clarté pour que d'autres puissent reproduire les conditions d'administration dans lesquelles les données relatives à la fiabilité, à la validité et (le cas échéant) aux normes ont été obtenues. Les variations autorisées dans les procédures d'administration doivent être clairement décrites. Le processus d'examen des demandes de variations supplémentaires des tests doit également être documenté.
- SEPT 4.16 : Les instructions présentées aux candidats doivent être suffisamment détaillées pour qu'ils puissent répondre à une tâche de la manière prévue par le concepteur du test. Le cas échéant, des échantillons de matériel, des questions d'entraînement ou des exemples de questions, des critères de notation et un item représentatif identifié avec chaque format d'item ou domaine majeur de la classification ou du domaine du test doivent être fournis aux candidats avant l'administration du test, ou doivent être inclus dans le matériel de test en tant que partie intégrante des instructions d'administration standard.
- SEPT 4.18 : Les procédures de notation et, le cas échéant, les critères de notation, doivent être présentés par le concepteur du test avec suffisamment de détails et de clarté pour maximiser la précision de la notation. Les instructions relatives à l'utilisation des échelles d'évaluation ou au calcul des scores obtenus par codage, échelonnement ou classification des réponses construites doivent être claires. Ceci est particulièrement important pour les questions à réponse étendue telles que les tâches de performance, les portfolios et les essais.
- SEPT 4.20 : Le processus de sélection, de formation, de qualification et de contrôle des correcteurs doit être spécifié par le concepteur du test. Le matériel de formation, tel que les grilles de notation et les exemples de réponses des candidats qui illustrent les niveaux de l'échelle de notation, ainsi que les procédures de formation des correcteurs doivent permettre d'atteindre un degré de précision et d'accord entre les correcteurs qui

permette d'interpréter les notes comme le concepteur du test l'avait initialement prévu. Les spécifications doivent également décrire les processus d'évaluation de la cohérence des évaluateurs et de la dérive potentielle dans le temps de la notation des évaluateurs.

- SEPT 1.9 : Lorsqu'une validation repose en partie sur les opinions ou les décisions de juges, d'observateurs ou d'évaluateurs experts, les procédures de sélection de ces experts et d'obtention de jugements ou d'évaluations doivent être décrites en détail. Les qualifications et l'expérience des juges doivent être présentées. La description des procédures doit inclure la formation et les instructions fournies, indiquer si les participants sont parvenus à leurs décisions de manière indépendante et signaler le niveau d'accord atteint. Si les participants ont interagi les uns avec les autres ou échangé des informations, les procédures par lesquelles ils ont pu s'influencer mutuellement doivent être exposées.
- SEPT 3.4 : Les participants aux tests doivent bénéficier d'un traitement comparable pendant l'administration et la notation des tests.
- SEPT 4.21 : Lorsque les utilisateurs du test sont responsables de la notation et que la notation nécessite le jugement du correcteur, l'utilisateur du test a la responsabilité de fournir une formation et des instructions adéquates aux correcteurs et d'examiner l'accord et la précision des correcteurs. Le concepteur du test doit documenter le niveau attendu de concordance et de précision des correcteurs et doit fournir autant de conseils techniques que possible pour aider les utilisateurs du test à satisfaire à cette norme.

4. Données pour la validation

- SEPT 1.8 : La composition de tout échantillon de participants à un test à partir duquel des preuves de validité sont obtenues doit être décrite de manière aussi détaillée que possible et pratique, y compris les principales caractéristiques sociodémographiques et développementales pertinentes.
- SEPT 1.13 : Preuves concernant la structure interne (c) : « Si la justification de l'interprétation des résultats d'un test pour une utilisation donnée dépend de prémisses concernant les relations entre les éléments du test ou entre les parties du test, des preuves concernant la structure interne du test doivent être fournies.
- SEPT 1.15 : Preuves concernant la structure interne (c) : « Lorsqu'il est suggéré d'interpréter les performances sur des items spécifiques ou de petits sous-ensembles d'items, il convient de fournir la justification et les preuves pertinentes à l'appui de cette interprétation. Lorsque l'interprétation des réponses à des items individuels est probable, mais n'est pas recommandée par le développeur, l'utilisateur doit être averti qu'il ne doit pas procéder à de telles interprétations.
- SEPT 4.7 : Les procédures utilisées pour élaborer, réviser et tester les items et pour sélectionner les éléments dans la réserve d'items doivent être documentées.
- SEPT 4.10 : Lorsqu'un concepteur de tests évalue les propriétés psychométriques des items, le modèle utilisé à cette fin (par exemple, la théorie classique des tests, la théorie

de la réponse à l'item ou un autre modèle) doit être documenté. L'échantillon utilisé pour estimer les propriétés des items doit être décrit et doit être d'une taille et d'une diversité adéquates pour la procédure. Le processus de sélection des items et les données utilisées pour la sélection, telles que la difficulté de l'item, la discrimination de l'item ou le fonctionnement différentiel de l'item (DIF) pour les principaux groupes d'examineurs, doivent également être documentés. Lorsque des méthodes basées sur des modèles (par exemple, la TRI) sont utilisées pour estimer les paramètres des items lors de l'élaboration des tests, le modèle de réponse de l'item, les procédures d'estimation et les preuves de l'adéquation du modèle doivent être documentés.

- SEPT 4.14 : Pour un test dont la durée est limitée, la recherche sur le développement des tests doit examiner dans quelle mesure les scores incluent une composante de vitesse et doit évaluer la pertinence de cette composante, compte tenu du domaine que le test est censé mesurer.
- SEPT 5.1 : Les utilisateurs des tests doivent recevoir des explications claires sur les caractéristiques, la signification et l'interprétation prévue des scores de l'échelle, ainsi que sur leurs limites.

5. Variance non pertinente de la construction et équité des tests

- SEPT 3.1 : Les responsables de l'élaboration, de la révision et de l'administration des tests doivent concevoir toutes les étapes du processus de test de manière à promouvoir des interprétations valides des scores pour les utilisations prévues, et ce pour le plus grand nombre possible d'individus et de sous-groupes pertinents de la population visée.
- SEPT 3.2 : Les concepteurs de tests sont chargés de mettre au point des tests qui mesurent le concept visé et de réduire au minimum le risque que les tests soient influencés par des caractéristiques non pertinentes pour le concept, telles que des caractéristiques linguistiques, communicationnelles, cognitives, culturelles, physiques ou autres.
- SEPT 3.5 : Les concepteurs de tests doivent spécifier et documenter les dispositions qui ont été prises pour l'administration des tests et les procédures de notation afin d'éliminer les obstacles non pertinents pour la construction de tous les sous-groupes pertinents de la population des participants aux tests.
- SEPT 4.5 : Si le concepteur du test indique que les conditions d'administration peuvent varier d'un candidat ou d'un groupe à l'autre, les variations autorisées dans les conditions d'administration doivent être identifiées. Une justification de l'autorisation de conditions différentes et toute exigence relative à l'autorisation de conditions différentes doivent être documentées.

6. Analyse des items et des tests

- SEPT 2.0 : Des preuves appropriées de fiabilité/précision doivent être fournies pour l'interprétation de chaque utilisation prévue du score.
- SEPT 2.1 : L'intervalle de réplication sur lequel la fiabilité/précision est évaluée doit être clairement indiqué, ainsi qu'une justification du choix de cette définition, compte tenu de la situation dans laquelle le test a été administré.
- SEPT 2.2 : Les preuves fournies pour la fiabilité/précision des scores doivent être cohérentes avec le domaine de réplication associé aux procédures de test, et avec les interprétations prévues pour l'utilisation des scores du test.
- SEPT 2.3 : Pour chaque score total, sous-score ou combinaison de scores à interpréter, les estimations des indices pertinents de fiabilité/précision doivent être indiquées.
- SEPT 2.4 : Lorsque l'interprétation des résultats d'un test met l'accent sur les différences entre deux scores observés d'un individu ou deux moyennes d'un groupe, les données de fiabilité/précision, y compris les erreurs standard, doivent être fournies pour ces différences.
- SEPT 2.5 : Les procédures d'estimation de la fiabilité doivent être cohérentes avec la structure du test.
- SEPT 2.7 : Lorsqu'un jugement subjectif intervient dans la notation d'un test, il convient de fournir des preuves de la cohérence entre les évaluateurs et de la cohérence au sein de l'examen pour des mesures répétées. Une distinction claire doit être faite entre les données de fiabilité basées sur (a) des panels indépendants d'évaluateurs notant les mêmes performances ou produits, (b) un seul panel notant des performances successives ou de nouveaux produits, et (c) des panels indépendants notant des performances successives ou de nouveaux produits.
- SEPT 2.10 : Lorsque des variations significatives sont autorisées dans les tests ou les procédures d'administration des tests, des analyses de fiabilité/précision distinctes doivent être fournies pour les scores obtenus dans le cadre de chaque variation majeure si des échantillons de taille adéquate sont disponibles.
- SEPT 2.13 : L'erreur standard de mesure, à la fois globale et conditionnelle (si elle est rapportée), doit être fournie en unités de chaque score rapporté.
- SEPT 2.14 : Lorsque cela est possible et approprié, les erreurs standard de mesure conditionnelles doivent être rapportées à plusieurs niveaux de score, à moins qu'il ne soit prouvé que l'erreur standard est constante entre les niveaux de score. Lorsque des seuils sont spécifiés pour la sélection ou la classification, les erreurs standard de mesure doivent être indiquées à proximité de chaque seuil.
- SEPT 2.16 : Lorsqu'un test ou une combinaison de mesures est utilisé pour prendre une décision de classification, il convient de fournir des estimations du pourcentage de participants au test qui seraient classés de la même manière lors de deux répétitions de la procédure.

- SEPT 2.19 : Chaque méthode de quantification de la fiabilité/précision des résultats doit être décrite clairement et exprimée en termes de statistiques appropriées à la méthode. Les procédures d'échantillonnage utilisées pour sélectionner les participants aux tests pour les analyses de fiabilité/précision et les statistiques descriptives sur ces échantillons, sous réserve des obligations en matière de protection de la vie privée le cas échéant, doivent être indiquées.
- SEPT 2.20 : Si les coefficients de fiabilité sont ajustés pour tenir compte de la restriction de l'intervalle ou de la variabilité, la procédure d'ajustement et les coefficients ajustés et non ajustés doivent être indiqués. Les écarts-types du groupe réellement testé et de la population cible, ainsi que la justification de l'ajustement, doivent être présentés.
- SEPT 3.3 : Les responsables de l'élaboration des tests doivent inclure les sous-groupes concernés dans les études de validité, de fiabilité/précision et autres études préliminaires utilisées lors de l'élaboration du test.
- SEPT 3.6 : Lorsque des preuves crédibles indiquent que les scores d'un test peuvent différer dans leur signification pour des sous-groupes pertinents de la population de candidats prévue, les développeurs et/ou les utilisateurs du test sont responsables de l'examen des preuves de la validité des interprétations des scores pour les utilisations prévues pour les individus de ces sous-groupes. Ce qui constitue une différence significative dans les résultats des sous-groupes et les mesures prises en réponse à de telles différences peuvent être définis par les lois applicables.
- SEPT 4.10 : Lorsqu'un concepteur de tests évalue les propriétés psychométriques des items, le modèle utilisé à cette fin (par exemple, la théorie classique des tests, la théorie de la réponse à l'item ou un autre modèle) doit être documenté. L'échantillon utilisé pour estimer les propriétés des items doit être décrit et doit être d'une taille et d'une diversité adéquates pour la procédure. Le processus de sélection des items et les données utilisées pour la sélection, telles que la difficulté de l'item, la discrimination de l'item ou le fonctionnement différentiel de l'item (FDI-DIF) pour les principaux groupes de candidats, doivent également être documentés. Lorsque des méthodes basées sur un modèle (par exemple, la TRI) sont utilisées pour estimer les paramètres des items lors de l'élaboration du test, le modèle de réponse de l'item, les procédures d'estimation et les preuves de l'adéquation du modèle doivent être documentés.

7. Développement d'items et 8. Développement de l'examen

- SEPT 4.7 : Les procédures utilisées pour élaborer, réviser et tester les éléments et pour sélectionner les éléments dans la banque de questions doivent être documentées.
- SEPT 4.8 : Le processus de révision des tests doit inclure des analyses empiriques et/ou le recours à des juges experts pour réviser les items et les critères de notation. Lorsque des juges experts sont utilisés, leurs qualifications, expériences pertinentes et caractéristiques démographiques doivent être documentées, de même que les instructions et la formation au processus d'examen des items que les juges reçoivent.

- SEPT 4.9 : Lorsque des essais d'items ou de formulaires de test sont effectués, les procédures utilisées pour sélectionner le(s) échantillon(s) de participants au test ainsi que les caractéristiques résultantes de l'(des) échantillon(s) doivent être documentées. L'échantillon doit être aussi représentatif que possible de la population à laquelle le test est destiné.
- SEPT 7.4 : La documentation relative au test doit résumer les procédures d'élaboration du test, y compris les descriptions et les résultats des analyses statistiques utilisées pour l'élaboration du test, les preuves de la fiabilité/précision des scores et la validité des interprétations recommandées, ainsi que les méthodes d'établissement des scores de coupure de la performance.

9. Administration de l'examen et 10. Accommodement pour un examen

- SEPT 4.5 : Si le concepteur du test indique que les conditions d'administration peuvent varier d'un candidat ou d'un groupe à l'autre, les variations autorisées dans les conditions d'administration doivent être identifiées. Une justification de l'autorisation de conditions différentes et toute exigence relative à l'autorisation de conditions différentes doivent être documentées.
- SEPT 4.15 : Les instructions relatives à l'administration du test doivent être présentées avec suffisamment de clarté pour que d'autres puissent reproduire les conditions d'administration dans lesquelles les données relatives à la fiabilité, à la validité et (le cas échéant) aux normes ont été obtenues. Les variations autorisées dans les procédures d'administration doivent être clairement décrites. Le processus d'examen des demandes de variations supplémentaires des tests doit également être documenté.
- SEPT 4.16 : Les instructions présentées aux candidats doivent être suffisamment détaillées pour qu'ils puissent répondre à une tâche de la manière prévue par le concepteur du test. Le cas échéant, des échantillons de matériel, des questions d'entraînement ou des exemples de questions, des critères de notation et un item représentatif identifié avec chaque format d'item ou domaine majeur de la classification ou du domaine du test doivent être fournis aux candidats avant l'administration du test, ou doivent être inclus dans le matériel de test en tant que partie intégrante des instructions d'administration standard.
- SEPT 6.0 : Pour permettre une interprétation utile des résultats, les instruments d'évaluation doivent être assortis de procédures établies pour l'administration, la notation, la communication et l'interprétation des tests. Les personnes chargées de l'administration, de la notation, de la communication et de l'interprétation doivent bénéficier d'une formation et d'un soutien suffisants pour les aider à suivre les procédures établies. Le respect des procédures établies doit être contrôlé et toute erreur matérielle doit être documentée et, si possible, corrigée.
- SEPT 6.1 : Les administrateurs de tests doivent suivre attentivement les procédures standardisées d'administration et de notation spécifiées par le concepteur du test et les instructions éventuelles de l'utilisateur du test.

- SEPT 6.7 : Les utilisateurs de tests ont la responsabilité de protéger à tout moment la sécurité du matériel de test.
- SEPT 7.0 : Les informations relatives aux tests doivent être clairement documentées afin que ceux qui utilisent les tests puissent prendre des décisions éclairées concernant le test à utiliser dans un but spécifique, la manière d'administrer le test choisi et la manière d'interpréter les résultats du test.

11. Établissement d'une note de passage

- SEPT 5.21 : Lorsque les interprétations de score proposées impliquent une ou plusieurs notes de passage (*cut score*) la justification et les procédures utilisées pour établir ces notes doivent être clairement documentées.
- SEPT 5.17 : Lorsque les scores de tests qui ne peuvent pas être mis en équivalence sont liés, des preuves directes de la comparabilité des scores doivent être fournies, et la population de candidats pour laquelle la comparabilité des scores s'applique doit être clairement spécifiée. La justification spécifique et les preuves requises dépendront en partie des utilisations prévues pour lesquelles la comparabilité des scores est revendiquée.

12. Notation et rapports

- SEPT 4.18 : Les procédures de notation et, le cas échéant, les critères de notation doivent être présentés par le concepteur du test avec suffisamment de détails et de clarté pour maximiser la précision de la notation. Les instructions relatives à l'utilisation des échelles d'évaluation ou au calcul des scores obtenus par codage, échelonnement ou classification des réponses construites doivent être claires. Ceci est particulièrement important pour les questions à réponse étendue telles que les tâches de performance, les portfolios et les essais.
- SEPT 4.20 : Le processus de sélection, de formation, de qualification et de contrôle des correcteurs doit être spécifié par le concepteur du test. Le matériel de formation, tel que les grilles de notation et les exemples de réponses des candidats qui illustrent les niveaux de l'échelle de notation, et les procédures de formation des correcteurs doivent permettre d'obtenir un degré de précision et d'accord entre les correcteurs qui permette d'interpréter les notes comme le concepteur du test l'avait initialement prévu. Les spécifications doivent également décrire les processus d'évaluation de la cohérence des évaluateurs et de la dérive potentielle dans le temps de la notation des évaluateurs.
- SEPT 4.21 : Lorsque les utilisateurs du test sont responsables de la notation et que la notation nécessite le jugement du correcteur, l'utilisateur du test a la responsabilité de fournir une formation et des instructions adéquates aux correcteurs et d'examiner l'accord et la précision des correcteurs. Le concepteur du test doit documenter le niveau attendu de concordance et de précision des correcteurs et doit fournir autant de conseils techniques que possible pour aider les utilisateurs du test à satisfaire à cette norme.

- SEPT 4.22 : Les concepteurs de tests doivent préciser les procédures utilisées pour interpréter les résultats des tests et, le cas échéant, les échantillons normatifs ou de normalisation ou le critère utilisé.
- SEPT 5.0 : Les résultats des tests doivent être obtenus d'une manière qui permette de les interpréter en fonction de l'utilisation qui en est faite. Les concepteurs et les utilisateurs de tests doivent documenter les preuves de l'équité, de la fiabilité et de la validité des résultats des tests pour l'utilisation proposée.
- SEPT 5.1 : Les utilisateurs des tests doivent recevoir des explications claires sur les caractéristiques, la signification et l'interprétation prévue des scores de l'échelle, ainsi que sur leurs limites.
- SEPT 5.2 : Les procédures d'élaboration des échelles utilisées pour l'établissement des scores et la justification de ces procédures doivent être décrites clairement.
- SEPT 5.3 : S'il y a de bonnes raisons de croire que des interprétations erronées d'une échelle de score sont possibles, les utilisateurs du test doivent être explicitement avertis.
- SEPT 5.8 : Les normes, si elles sont utilisées, doivent se référer à des populations clairement décrites. Ces populations doivent comprendre des individus ou des groupes avec lesquels les utilisateurs des tests souhaitent habituellement comparer leurs propres candidats.
- SEPT 6.8 : Les responsables de la notation des tests doivent établir des protocoles de notation. La notation des tests impliquant un jugement humain doit inclure des rubriques, des procédures et des critères de notation. Lorsque la notation de réponses complexes est effectuée par ordinateur, la précision de l'algorithme et des processus doit être documentée.
- SEPT 6.9 : Les responsables de la notation des tests doivent établir et documenter les processus et les critères de contrôle de la qualité. Une formation adéquate doit être dispensée. La qualité de la notation doit être contrôlée et documentée. Toute source systématique d'erreurs de notation doit être documentée et corrigée.
- SEPT 6.10 : Lorsque les résultats d'un test sont publiés, les responsables des programmes de test doivent fournir des interprétations adaptées au public visé. Les interprétations doivent décrire dans un langage simple ce que le test couvre, ce que les scores représentent, la précision/fiabilité des scores, et comment les scores sont destinés à être utilisés.
- SEPT 6.12 : Lorsque des informations au niveau du groupe sont obtenues en agrégeant les résultats de tests partiels effectués par des individus, les preuves de validité et de fiabilité/précision doivent être rapportées pour le niveau d'agrégation auquel les résultats sont rapportés. Les scores ne doivent pas être rapportés pour des individus sans preuves appropriées pour soutenir les interprétations pour les utilisations prévues.
- SEPT 7.3 : Lorsque l'information est disponible et partagée de manière appropriée, les documents relatifs aux tests doivent citer un ensemble représentatif d'études relatives aux utilisations générales et spécifiques du test.

- SEPT 7.5 : Les documents relatifs aux tests doivent indiquer les caractéristiques pertinentes des personnes ou des groupes de personnes qui ont participé à la collecte des données liées à l'élaboration ou à la validation des tests (par exemple, informations démographiques, situation professionnelle, niveau scolaire) ; la nature des données fournies (par exemple, données prédictives, données de critères) ; la nature des jugements portés par les experts en la matière (par exemple, liens avec la validation du contenu) ; les instructions fournies aux participants à la collecte des données pour leurs tâches spécifiques ; et les conditions dans lesquelles les données du test ont été collectées dans le cadre de l'étude de validité.
- SEPT 7.8 : La documentation du test doit comprendre des instructions détaillées sur la manière dont le test doit être administré et noté.
- SEPT 7.13 : Les documents d'appui (par exemple, les manuels de test, les manuels techniques, les guides de l'utilisateur et les documents complémentaires) doivent être mis à la disposition des personnes concernées en temps utile.

