

Université de Sherbrooke

**Exploration de la validité en tant qu'impératif social
dans le contexte de l'évaluation des apprentissages
en pédagogie des sciences de la santé**

Par
Mélanie Marceau
Programme de sciences cliniques

Thèse présentée à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade philosophiae doctor (Ph.D.)
en sciences cliniques (pédagogie des sciences de la santé)

Sherbrooke, Québec, Canada
Novembre, 2019

Membres du jury d'évaluation

Pre Annie Carrier	Présidente du jury, programmes Recherche en sciences de la santé, Université de Sherbrooke
Pre Frances Gallagher	Directrice, programmes Recherche en sciences de la santé, Université de Sherbrooke
Pre Christina St-Onge	Directrice, programmes Recherche en sciences de la santé, Université de Sherbrooke
Pre Meredith Young	Directrice, Département de médecine, Université McGill
Pr Pascal Ndinga	Évaluateur externe au programme, Département d'éducation et pédagogie, Université du Québec à Montréal
Pre Johanne Gagnon	Évaluatrice externe à l'Université, Faculté des sciences infirmières, Université Laval

©Mélanie Marceau, 2019

SOMMAIRE

**La validité en tant qu'impératif social dans le contexte de l'évaluation des
apprentissages en pédagogie des sciences de la santé**

Par

Mélanie Marceau

Programme de sciences cliniques (pédagogie des sciences de la santé)

Thèse présentée à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du diplôme de philosophiae doctor (Ph.D.) en sciences cliniques,
Faculté de médecine et des sciences de la santé,
Université de Sherbrooke, Québec, Canada, J1H 5N4

Introduction: Les processus de validation existants semblent inadaptés aux pratiques évaluatives intégrées dans les programmes professionnalisants en sciences de la santé. L'insatisfaction des chercheurs et du personnel enseignant envers les processus de validation a contribué à l'émergence de la conceptualisation de la validité en tant qu'impératif social.

But et objectifs : Explorer le concept de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages, soit : 1) décrire le concept de la validité en tant qu'impératif social; 2) explorer l'acceptabilité et la faisabilité anticipée de l'application du concept et 3) décrire les éléments à considérer lors de l'application du concept.

Méthode et résultats : Une recherche exploratoire en trois phases, fondée sur un cadre paradigmatique constructiviste, a été utilisée. *Phase 1* : une analyse de concept a permis de décrire les trois antécédents, les quatre attributs et le conséquent du concept émergent. *Phase 2* : une approche descriptive interprétative a servi à explorer l'acceptabilité et la faisabilité anticipée du concept auprès 1) du personnel enseignant et de membres de comités d'évaluation des apprentissages (n = 23) et 2) de personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada (n = 7) à l'aide de groupes de discussion focalisée (n = 5) et d'entrevues individuelles (n = 11). Les participants provenant des deux profils ont exprimé le besoin d'obtenir des clarifications à propos de termes (p.ex. société) et ont nommé des défis envisagés lors de l'application du concept. Parmi les différences observées entre les deux profils, les membres du personnel enseignant ont mentionné avoir réalisé l'importance de leur rôle au sujet de la validité alors que les personnes engagées en recherche au Canada ont nommé l'importance de rendre explicite le rôle social de la validité. *Phase 3* : une approche descriptive interprétative a été utilisée pour intégrer les données recueillies lors des phases 1 et 2 et décrire les éléments à considérer lors de l'application du concept. Quatre éléments sont proposés pour inclure l'impératif social dans un processus de validation, soit : 1) planification de l'évaluation, 2) planification de l'administration de l'évaluation, 3) documentation des preuves de validité selon un plan structuré et flexible et 4) planification de l'interprétation de l'ensemble des résultats.

Conclusion : Le concept de validité en tant qu'impératif social s'avère une piste de solution prometteuse pour combler l'écart entre les pratiques évaluatives et le processus de validation. Les balises élaborées permettront aux personnes responsables de l'évaluation de sélectionner les éléments pertinents à leur contexte. En facilitant le travail des personnes responsables de l'évaluation, nous espérons améliorer la formation des futurs professionnels de la santé.

Mots-clés : Évaluation des apprentissages, validité, impératif social

SUMMARY

Validity as a social imperative in the context of assessment in health professional education

By

Mélanie Marceau

Clinical sciences Program (Health Professional Education)

A thesis presented to the Faculty of Medicine and Health Sciences in partial fulfillment of the requirements of the degree of Doctor of Philosophy in Clinical sciences,
Faculty of Medicine and Health Sciences,
Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

Introduction: Current validation processes appear to be insufficient for assessment practices used in health professions education programs. Researchers' and educators' dissatisfaction with validation processes has contributed to the emergence of a new conceptualization of validity, that is, validity as a social imperative.

Aim and objectives: Explore the concept of validity as a social imperative in the context of assessment. The objectives are 1) to describe the emerging concept of validity as a social imperative; 2) explore the acceptability and anticipated feasibility of the concept for use and 3) describe the elements to be considered when applying the concept.

Method and results: A three phase exploratory research design aligned with a constructivist paradigm was used. *Phase 1:* A concept analysis described the three antecedents, four attributes and one consequent of the emerging concept. *Phase 2:* A qualitative interpretive description approach was used to explore the acceptability and anticipated feasibility of the concept with 1) faculty and assessment committee members (n = 23) and 2) individuals engaged in health science education research in Canada (n = 7) using focus groups (n = 5) and individual interviews (n = 11). Both stakeholder groups expressed the need for the clarification of certain terms (e.g. society) and anticipated challenges in applying the concept. Among the differences observed between the two profiles, faculty and assessment committee members reported that they realized the importance of their role regarding validity while individuals engaged in research in Canada identified the importance of making the role of society in validity explicit. *Phase 3:* A qualitative interpretive description approach was used to transform the data collected during phases 1 and 2 and describe the elements to be considered when applying the concept. Four elements are proposed to integrate social aspects into a validation process: 1) planning of the assessment, 2) planning of the administration of the assessment, 3) documenting validity evidences according to a structured and flexible plan, and 4) planning of validity evidence and assessment score.

Conclusion: The concept of validity as a social imperative is a promising solution to bridge the gap between assessment practices and the current available validation processes. By facilitating the work of those responsible for assessment, we hope to improve the training of future health professionals.

Keywords: Assessment, validity, social imperative

TABLE DES MATIÈRES

Introduction.....	1
1. Problématique	2
1.1 Énoncé du problème.....	2
1.2 Pertinence de l'étude.....	5
1.3 But et objectifs de la recherche	6
2. Recension des écrits	7
2.1 Démarche de recherche documentaire	7
2.2 Pratiques évaluatives.....	9
2.2.1 Évaluation des apprentissages	10
2.2.2 Rétroaction	14
2.2.3 Évaluation programmatique	16
2.2.4 Approche par compétences.....	18
2.3 Processus de validation	26
2.3.1 Conceptualisations de la validité	26
2.3.2 Théorie unifiée de la validité.....	28
2.3.3 Validation fondée sur l'accumulation des preuves.....	31
2.3.4 Critiques des théories modernes.....	34
2.4 Écart entre les pratiques évaluatives et les processus de validation existants	35
3. Méthode.....	38
3.1 Devis	38
3.2 Cadre paradigmatique	39
3.3 Phase 1 : Description du concept émergent	40
3.4 Phase 2 : Acceptabilité et faisabilité anticipée du concept	42
3.4.1 Analyse des données.....	47
3.4.2 Considérations éthiques.....	50
3.5 Phase 3 : Éléments à considérer lors de la mise en œuvre du concept de validité en tant qu'impératif social	51
3.5.1 Intégration des données	51
4. Méthode, résultats et discussion de la Phase 1 - Analyse de concept	55
Article 1.....	56
5. Méthode, résultats et discussion de la phase 2 – Acceptabilité et faisabilité anticipée ...	91
5.1 Article 2.....	92
5.2 Compléments méthodologiques.....	120
5.2.1 Personnel enseignant et membres de comités d'évaluation des apprentissages.....	120
5.2.2 Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada	123
5.3 Caractéristiques sociodémographiques complémentaires.....	125
5.3.1 Personnel enseignant et membres de comités d'évaluation des apprentissages.....	126

5.3.2 Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada	128
6. Méthode, résultats et discussion de la Phase 3 – Éléments à considérer	130
6.1 Article 3.....	131
6.2 Résultats complémentaires.....	153
7. Discussion générale.....	157
7.1 Principaux constats	158
7.1.1 Acceptabilité du concept de validité en tant qu'impératif social.....	158
7.1.2 Lien entre l'impératif social et la responsabilité sociale	159
7.1.3 Réponse aux préoccupations concernant les limites de processus de validation courants.....	160
7.1.4 Recours aux critères de rigueur des devis mixtes.....	161
7.1.5 Imputabilité du processus de validation	163
7.2 Forces et limites méthodologiques.....	165
7.2.1 Recherche exploratoire en trois phases	165
7.2.2 Analyse de concept.....	166
7.2.3 Approche descriptive interprétative	167
7.2.4 Collecte des données	167
7.2.5 Rigueur scientifique.....	169
7.3 Retombées.....	171
7.3.1 Retombées pour la pédagogie en sciences de la santé.....	171
7.3.2 Retombées pour la recherche.....	172
Conclusion	174
Références.....	175
ANNEXES.....	193
Annexe 1. Approbation éthique de l'Université de Sherbrooke	194
Annexe 2. Approbation éthique de l'Université McGill.....	195
Annexe 3. Approbation éthique de l'Université Laval	197
Annexe 4. Approbation éthique de l'Université de Montréal.....	198
Annexe 5. Invitation à participer Personnel enseignant et membres de comités d'évaluation des apprentissages.....	199
Annexe 6. Questionnaire sociodémographique Personnel enseignant et membres de comités d'évaluation des apprentissages.....	201
Annexe 7. Guide pour les entrevues individuelles Personnel enseignant et membres de comités d'évaluation des apprentissages.....	203
Annexe 8. Guide pour les groupes de discussion focalisée Personnel enseignant et membres de comités d'évaluation des apprentissages	210
Annexe 9. Résumé de l'analyse de discours (version française)	217
Annexe 10. Résumé de l'analyse de concept (version française).....	219
Annexe 11. Formulaire de consentement pour les entrevues individuelles Personnel enseignant et membres de comités d'évaluation des apprentissages	221
Annexe 12. Formulaire de consentement pour les groupes de discussion focalisée Personnel enseignant et membres de comités d'évaluation des apprentissages	224

Annexe 13. Invitation à participer Personnes engagées dans la recherche en pédagogie des sciences de la santé	227
Annexe 14. Questionnaire sociodémographique Personnes engagées dans la recherche en pédagogie des sciences de la santé	230
Annexe 15. Formulaire de consentement Personnes engagées dans la recherche en pédagogie des sciences de la santé.....	232
Annexe 16. Guide pour les entrevues individuelles Personnes engagées dans la recherche en pédagogie des sciences de la santé	235
Annexe 17. Résumé de l'analyse de discours (version anglaise)	240

LISTE DES TABLEAUX

Tableaux de la thèse

Tableau 1 Mots-clés utilisés pour la recension des écrits	8
Tableau 2 Conceptualisations de la validité	28
Tableau 3 Théorie unifiée de la validité	30
Tableau 4 Validation fondée sur l'accumulation des preuves	33
Tableau 5 Analyse de concept selon Rodgers	42
Tableau 6 Résumé des décisions méthodologiques de la phase 2	46
Tableau 7 Codes prédéterminés et exemples de codes émergents.....	48
Tableau 8 Extrait d'une matrice conceptuelle portant sur l'attribut 1 (phase 2).....	49
Tableau 9 Extrait d'une matrice conceptuelle (phase 3).....	53
Tableau 10 Caractéristiques sociodémographiques du personnel enseignant et des membres de comité d'évaluation des apprentissages	126
Tableau 11 Caractéristiques sociodémographiques des personnes engagées dans la recherche au Canada	129
Tableau 12 Description de l'expérience du participant.....	129
Tableau 13 Évolution du concept de validité en tant qu'impératif social	155

Tableaux de l'article 1

Table 1 Steps in Rodgers' method.....	64
Table 2 Categories of documents included in the analysis.....	69
Table 3 Disciplines represented in included documents.....	69

Tableaux de l'article 2

Table 1 Characteristics of the concept of validity as a social imperative.....	98
Table 2 Procedure used with stakeholders.....	99
Table 3 Sociodemographic characteristics of participants.....	104
Table 4 Users and leaders' perspectives	104

Tableau de l'article 3

Table 1 Attributes and guideposts to consider in validity as a social imperative.....	139
---	-----

LISTE DES FIGURES

Figures de la thèse

Figure 1 Diagramme de flux de la recension des écrits	9
Figure 2 Schématisation du processus de recherche.....	39
Figure 3 Contribution des participants.....	127
Figure 4 Méthodes d'évaluation des apprentissages utilisées par les participants	128
Figure 5 Simplification des attributs.....	153
Figure 6 Organisation des attributs.....	156

Figures de l'article 1

Figure 1 Document selection process	66
Figure 2 Results of the concept analysis of validity as a social imperative for assessment in health professional education.....	70

LISTE DES ABRÉVIATIONS

AERA	<i>American Educational Research Association</i>
APA	<i>American Psychological Association</i>
APC	Approche par compétences
ECOS	Examen clinique objectif structuré
EOS	Examen oral structuré
EPA	<i>Entrustable professional activity</i>
IVC	Indice de validité de contenu
NCME	<i>National Council on Measurement in Education</i>
SMUS	Société des médecins de l'Université de Sherbrooke
QRC	Questionnaire à réponses courtes
QCM	Questionnaire à choix multiples

À mon père Denis

REMERCIEMENTS

Cette thèse constitue un accomplissement qui se serait avéré impossible sans le soutien de plusieurs personnes formidables.

Tout d'abord, je remercie mes directrices de recherche : Pre Frances Gallagher, Pre Christina St-Onge et Pre Meredith Young. Vos forces complémentaires ont été un ingrédient essentiel à la réussite de mon doctorat. Vous avez célébré avec moi les bons coups et m'avez soutenue dans les moments les plus difficiles. À vos côtés, j'ai beaucoup grandi comme chercheuse, mais également comme personne.

Je remercie mes collègues et amies de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke et du Centre de pédagogie des sciences de la santé. Élise, Linda, Josée, Kathleen, votre écoute et votre complicité ont grandement contribué à ma persévérance et à mon succès.

Bien entendu, le financement demeure indispensable afin de me permettre de me consacrer pleinement aux travaux de doctorat. À cet égard, je suis reconnaissante au Programme MEES-Universités ainsi qu'à la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke.

Merci à mon amoureux, à mes parents, mes sœurs, mon beau-frère, mes neveux et nièces pour leur encouragement et leur compréhension. Merci également à mes amis de l'urgence, mes collègues de l'ÉSI et mes amis qui sont présents depuis le début de l'aventure. Vous m'avez permis de conserver un équilibre malgré les nombreuses heures de travail et les difficultés rencontrées.

INTRODUCTION

La validité en tant qu'impératif social est une conceptualisation émergente dans la littérature en pédagogie des sciences de la santé. Cette conceptualisation rend explicite la responsabilité des personnes (p.ex. les administrateurs de programme de formation et le personnel enseignant) et des instances (p.ex. ordres professionnels ou établissements d'enseignements) envers la qualité de l'évaluation des apprentissages des futurs professionnels de la santé afin de minimiser les conséquences sur l'apprenant et sur la société. En raison de son émergence, le projet doctoral a pour but d'explorer ce qu'est la conceptualisation de la validité en tant qu'impératif social. En cohérence avec la nature exploratoire de la recherche, le paradigme constructiviste a guidé les décisions méthodologiques.

La thèse est réalisée dans le cadre du programme en sciences cliniques, avec une spécialisation en pédagogie des sciences de la santé. Pour répondre aux objectifs de recherche, nous avons mené une recherche exploratoire en trois phases, fondée sur un cadre paradigmatique constructiviste. La présente thèse est divisée de manière à faciliter la compréhension de chacune des phases de la recherche. Tout d'abord, au chapitre 2, nous décrirons l'état actuel des connaissances découlant d'une recension des écrits. Nous présenterons ensuite la vue d'ensemble du processus de recherche au chapitre 3. Aux chapitres 4, 5 et 6, nous détaillerons la méthode, les résultats et la discussion spécifiques à chacune des trois phases de la recherche. Dans le chapitre 4, nous expliquerons de façon plus approfondie l'objectif 1 qui visait à décrire le concept émergent de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages. Dans le chapitre 5, nous décrirons le deuxième objectif de la recherche qui était d'explorer l'acceptabilité et la faisabilité anticipée du concept de validité en tant qu'impératif social. Dans le chapitre 6, nous précisons le troisième objectif qui visait à décrire les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social. Chacun des chapitres 4, 5 et 6 est constitué d'un article (article 1 pour la phase 1, article 2 pour la phase 2 et article 3 pour la phase 3) et de résultats complémentaires à ceux mentionnés dans l'article. Finalement, le chapitre 7 se veut une discussion générale portant sur les principaux constats de la présente recherche, les forces et limites de l'étude ainsi que les retombées de la recherche.

1. PROBLÉMATIQUE

Ce premier chapitre est composé de trois sections, lesquelles sont : l'énoncé du problème, la pertinence de l'étude ainsi que le but et les objectifs de la recherche.

1.1 Énoncé du problème

L'évaluation des apprentissages consiste à documenter le niveau de connaissance ou encore la performance d'un apprenant (Wass, van der Vleuten, Shatzer, & Jones, 2001) à l'aide de pratiques évaluatives. Une fois recueillies, ces données sont ensuite analysées et interprétées afin de poser une action ou de prendre une décision comme celle d'attribuer la note réussite ou échec de l'apprenant (Cook & Lineberry, 2016). Dans les programmes en sciences de la santé, les apprenants sont évalués à différents moments de leur parcours scolaire, que ce soit pour être admis dans un programme (Alexander & Cleland, 2018; Cook, Brydges, Ginsburg, & Hatala, 2015; Harden, 2015; Kane, 2013; Prideaux et al., 2011; Young, St-Onge, Xiao, Vachon Lachiver, & Torabi, 2018) pour vérifier l'acquisition de leurs compétences (Frank et al., 2010; Govaerts, 2015; Young et al., 2018) ou pour obtenir un diplôme ou un permis d'exercice (Hays, 2008). À tous ces niveaux, il est socialement impératif de pouvoir documenter la qualité de l'évaluation des apprentissages, c'est-à-dire de s'assurer que les décisions prises découlent d'une interprétation juste des résultats provenant de l'évaluation des apprentissages. Le terme « impératif » précise le rôle essentiel des personnes et instances impliquées dans l'évaluation des apprentissages. Celles-ci ont la responsabilité de poser des actions qui contribueront à la qualité de l'évaluation des apprentissages pour minimiser les conséquences sur l'apprenant ou la société.

Qu'elle soit à faible enjeu ou enjeu élevé, l'évaluation des apprentissages peut entraîner des conséquences positives ou négatives, qui peuvent être intentionnelles ou non, sur l'apprenant ou sur la société (Cook & Lineberry, 2016). Chez l'apprenant, l'évaluation des apprentissages peut occasionner, par exemple, de l'anxiété ou encore stimuler sa motivation lorsqu'il obtient de bons résultats scolaires (Cook & Lineberry, 2016). Pour la société, une

évaluation des apprentissages de qualité contribue à la sécurité du public en mesurant la compétence des professionnels (Cook & Lineberry, 2016).

Dans le contexte de l'évaluation des apprentissages, la validité correspond à la qualité de l'interprétation des résultats provenant de pratiques évaluatives. Les pratiques évaluatives ne peuvent être valides. Ce sont plutôt les interprétations ou les décisions issues des résultats des pratiques évaluatives qui sont jugées sur un continuum de plus à moins valide (Bertrand & Blais, 2004). Puisque l'interprétation des résultats découlant de l'évaluation des apprentissages engendre des conséquences, il s'avère primordial d'en vérifier la qualité (van der Vleuten & Schuwirth, 2005) afin de tirer des inférences justes sur les apprentissages (Cizek, Rosenberg, & Koon 2008; DeLuca, 2011; Downing, 2003; Messick, 1995). La qualité d'une évaluation est vérifiée à l'aide d'un processus de validation continu, qui implique l'utilisation de différentes stratégies de validation qui peuvent être quantitatives (analyses psychométriques) et qualitatives. Ces stratégies sont utilisées pour collecter les preuves de validité qui appuient l'interprétation des résultats et les décisions qui en découlent (Kane, 2006; Messick, 1995; Sireci & Sukin, 2013). Plus l'enjeu relié à l'évaluation est élevé (p. ex. un examen de certification), plus les preuves de validité soutenant les décisions (p. ex. réussite ou échec) doivent être documentées (Dijkstra, van der Vleuten, & Schuwirth, 2010; Fotheringham, 2010; Hubley & Zumbo, 2011; Kane, 2006; McAllister, Lincoln, Ferguson, & McAllister, 2013; Prideaux et al., 2011; Roberts, Newble, Jolly, Reed, & Hampton, 2006; Rosen, Salas, Silvestri, Wu, & Lazzara, 2008; Schuwirth & van der Vleuten, 2013; Scoles, Hawkins, & LaDuca, 2003; van der Vleuten & Schuwirth, 2005).

Afin de répondre aux besoins de la société (Arcand & Neumann, 2005; Hodges & Lingard, 2012; Royal College of Physicians and Surgeons of Canada, 2014) et aux normes des instances professionnelles visant à former des personnes compétentes, la majorité des établissements d'enseignement ont révisé leurs programmes de formation et ont intégré de nouvelles pratiques évaluatives (p. ex. évaluation programmatique et évaluation longitudinale). Toutefois, les stratégies de validation ont peu changé, ce qui provoque un écart entre les pratiques évaluatives et les processus de validation existants (Driessen, 2017; Hodges, 2015; Schuwirth & van der Vleuten, 2006; St-Onge, Young, Eva, & Hodges, 2017;

van der Vleuten & Dannefer, 2012; van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). Par exemple, les analyses psychométriques souvent utilisées dans les processus de validation se limitent à la mesure de la qualité d'une seule méthode d'évaluation, ce qui semble incompatible avec l'évaluation programmatique (van der Vleuten & Schuwirth, 2005).

Cet écart provoqué par un processus de validation inadapté aux pratiques évaluatives a mené à des questionnements que l'on retrouve dans les écrits scientifiques en pédagogie des sciences de la santé. Des chercheurs s'interrogent sur l'amélioration des processus de validation et la qualité des nouvelles pratiques évaluatives, en regardant trois angles différents, à savoir 1) comment sont appliqués les processus de validation (Cizek, Rosenberg, & Koons, 2008; Cook, Brydges, Zendejas, Hamstra, & Hatala, 2013; Cook, Zendejas, Hamstra, Hatala, & Brydges, 2014); 2) comment améliorer l'opérationnalisation des processus de validation existants (Cook et al., 2015; Cook & Beckman, 2006; Downing, 2003; Schuwirth & van der Vleuten, 2012) et 3) comment porter un regard différent sur le processus de validation. Ce troisième angle de recherche et de réflexions a mené les personnes engagées dans la recherche en pédagogie des sciences de la santé à suggérer une vision de la validité plus englobante, et à considérer davantage les conséquences de l'évaluation des apprentissages sur l'apprenant et sur la société (Cook & Lineberry, 2016).

Dans un examen des écrits scientifiques, les auteurs d'une analyse du discours (St-Onge et al., 2017) ont répertorié trois différentes conceptualisations de la validité, soit 1) la validité comme une caractéristique d'un test (la validité d'un outil est statique, ce qui permet de dire qu'un outil est valide); 2) la validité comme une accumulation de preuves (la validité est plutôt considérée comme évolutive, vue comme une quête qui nécessite plusieurs preuves de validité) et 3) la validité en tant qu'impératif social. Cette troisième conceptualisation, la validité en tant qu'impératif social, regroupe pour une première fois les réflexions portant sur une vision plus récente de la validité de l'évaluation des apprentissages. Les personnes ayant cette conceptualisation de la validité soulignent l'importance de la responsabilité sociale des personnes qui planifient, administrent et monitorent l'évaluation des apprentissages. La validité en tant qu'impératif social se caractérise par une considération marquée des

conséquences de l'évaluation non seulement pour l'apprenant, mais pour l'ensemble de la société. Ces considérations doivent être mises au premier plan pendant le développement de l'évaluation des apprentissages et pendant tout le processus de validation (St-Onge et al., 2017). La qualité de l'évaluation des apprentissages doit être vue de manière holistique, c'est-à-dire qu'elle doit être examinée dans son ensemble (p.ex. évaluation programmatique) et non examinée de façon individuelle (p.ex. méthode d'évaluation).

1.2 Pertinence de l'étude

Avec l'état actuel des connaissances et les nombreux questionnements dans les écrits scientifiques, un processus de validation qui intègre les principes de validité en tant qu'impératif social pourrait permettre une validation plus appropriée des pratiques évaluatives actuellement privilégiées dans les programmes professionnalisants en sciences de la santé. Par exemple, les processus de validation pourraient être mieux adaptés aux nouvelles pratiques évaluatives comme l'évaluation programmatique, notamment par la considération de l'ensemble des données recueillies à propos de la performance de l'apprenant.

Une meilleure validation des pratiques évaluatives est essentielle afin que les compétences des futurs professionnels de la santé soient évaluées correctement, minimisant ainsi les conséquences sur l'individu et sur la société. Bien que les auteurs d'une analyse du discours (St-Onge et al., 2017) aient identifié l'existence du concept de validité en tant qu'impératif social, nous avons une compréhension sommaire de ce concept en émergence. L'état actuel des connaissances ne nous permet pas de juger si l'utilisation du concept représente une solution aux limites des processus de validation existants. Une meilleure compréhension de cette conceptualisation de la validité permettrait d'en vérifier l'utilisation concrète dans les programmes de formation en sciences de la santé.

1.3 But et objectifs de la recherche

La présente étude vise à explorer le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé, auprès du personnel enseignant, des membres de comités d'évaluation des apprentissages et des chercheurs.

Les objectifs spécifiques sont les suivants :

- 1) décrire le concept émergent de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages;
- 2) explorer l'acceptabilité et la faisabilité anticipée du concept de validité en tant qu'impératif social;
- 3) décrire les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social.

2. RECENSION DES ÉCRITS

Nous avons structuré la recension des écrits en trois thèmes, soit : 1) les pratiques évaluatives utilisées dans les programmes en sciences de la santé; 2) les processus de validation existants et 3) l'écart entre les pratiques évaluatives et les processus de validation existants. Ces thèmes ont été sélectionnés pour mieux explorer et circonscrire les défis vécus par les personnes responsables de l'évaluation en pédagogie des sciences de la santé. Nous débuterons ce chapitre par une description de la démarche de recherche documentaire et présenterons ensuite l'état des connaissances pour chacun des thèmes identifiés.

2.1 Démarche de recherche documentaire

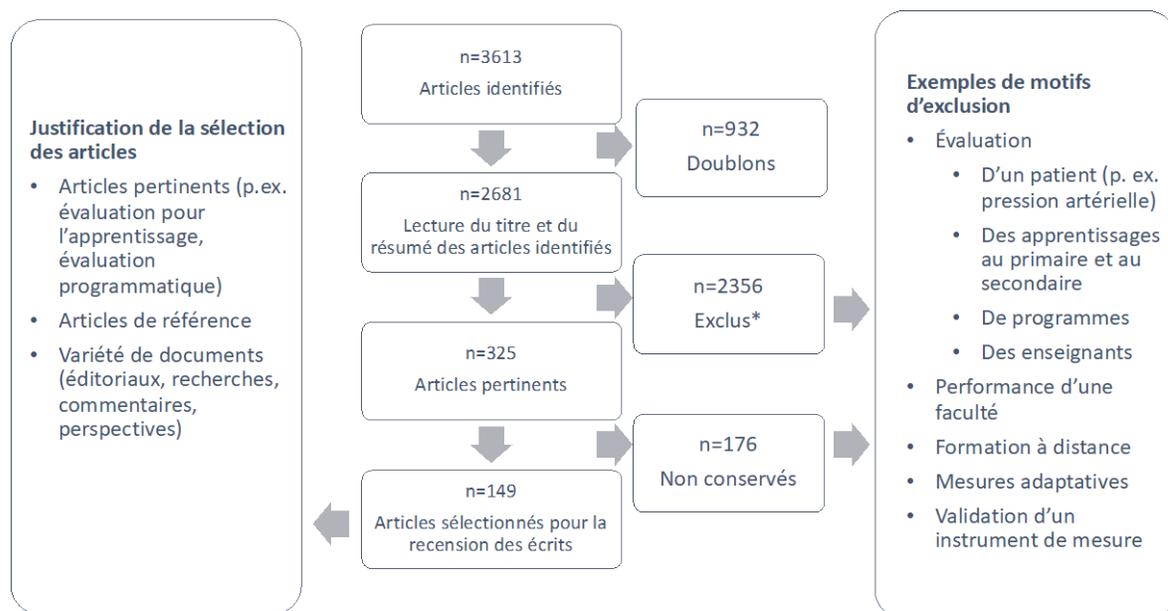
Pour effectuer une recension rigoureuse des écrits, nous avons travaillé en collaboration avec une bibliothécaire académique de la Faculté de médecine et des sciences de la santé. L'étudiante et ses directrices avaient préalablement identifié les mots-clés pertinents. La consultation avec la bibliothécaire a permis de peaufiner la stratégie de recherche et de bonifier le choix des banques de données. Nous avons ciblé les documents en langue française ou anglaise, publiés entre 1995 et 2018 et repérés dans les banques de données suivantes : PsychInfo, Medline, Academic Search Complete, Cinahl, Education Source et ERIC. Nous avons également consulté les banques de données Francis et Pascal pour obtenir des documents publiés entre 1995 et 2015. Pour ces deux banques de données, nous n'avons pas eu la possibilité de rechercher les documents entre 2015 et 2018 puisqu'elles ne sont plus mises à jour depuis 2015. Nous avons choisi l'année de publication 1995 pour baliser la période visée par la recherche documentaire. Cette date correspond à celle de la première publication de Messick (1995) sur la validité unifiée qui avait été révisée par les pairs et qui venait révolutionner la façon de conceptualiser la validité. Les mots-clés de la recherche documentaire figurent au tableau 1.

Tableau 1
Mots-clés utilisés pour la recension des écrits

Concept 1 Évaluation		Concept 2 Population		Concept 3 Validation
“Competency-based”		“Higher education”		Valid*
OR		OR		OR
“work* based” N5		“Undergraduate stud*”		Standard*
OR		OR		OR
“Outcome* based”		“Health profession*”		Test*
N5		OR		OR
OR	AND	“Medical education”	AND/OR	Qualit*
“Performance based” N5		OR		
OR		“Professional education”		
Assess*		OR		
OR		“Professional education”		
Evaluat*				
OR				
Test*				
OR				
Measure*				

Nous illustrons, dans le diagramme de flux de la recension des écrits (figure 1), le processus de sélection des documents identifiés dans le cadre de cette recension des écrits. Il peut sembler surprenant pour le lecteur de ne pas retrouver une recension des écrits classique qui présente une synthèse des recherches sur le phénomène étudié. Nous avons visé une grande variété de documents considérant le caractère émergent du concept à l'étude en sélectionnant des éditoriaux, des recherches, des commentaires ou encore des documents rédigés par différentes instances professionnelles. Nous avons également choisi des articles de référence cités par plusieurs auteurs et exclu les documents ne portant pas sur l'évaluation des apprentissages en sciences de la santé (p. ex. évaluation des besoins, évaluation de l'enseignant, évaluation d'un programme). Nous avons également exclu les recherches portant sur la validation de méthodes d'évaluation des apprentissages qui ne mesurent pas spécifiquement la performance d'un apprenant (p.ex. questionnaire à choix multiples) puisque les processus de validation existants ne sont pas problématiques. Nous avons conservé un nombre total de 149 documents.

Figure 1
Diagramme de flux de la recension des écrits



2.2 Pratiques évaluatives

Les programmes de formation en sciences de la santé s'appuient de plus en plus sur les philosophies et les principes de l'approche par compétences (APC) afin de préparer les futurs professionnels de la santé à jouer leur rôle dans la société (Axley, 2008; Frank et al., 2010; Frank, Snell, Englander, & Holmboe, 2017; Hodges & Lingard, 2012; Lucey, Thibault, & Ten Cate, 2018). L'implantation de l'APC dans les programmes en sciences de la santé s'effectue depuis le début des années 2000 et a nécessité un ajustement des pratiques évaluatives pour assurer la cohérence de l'ensemble du processus pédagogique (Baartman, Bastiaens, Kirschner, & van der Vleuten, 2007; Cizek, Kosh, & Toutkoushian, 2018; Dijkstra et al., 2010; Hawkins et al., 2015).

Nous utiliserons le terme « pratique évaluative » dans l'ensemble de la thèse. Nous nous sommes appuyés sur la définition précisée par Schuwirth et van der Vleuten (2019) : “*any purported and formal action to obtain information about the competence and performance of a candidate* (p. 277)”. Ainsi, les pratiques évaluatives englobent les méthodes d'évaluation (p. ex. questionnaire à choix multiples (QCM) ou examen oral structuré) ainsi que les visions et philosophies liées à l'évaluation (p. ex. évaluation pour l'apprentissage et évaluation

programmatische). Dans cette section de la recension des écrits, nous aborderons l'évaluation des apprentissages, la rétroaction, l'évaluation programmatique, l'approche par compétence (APC) ainsi que les pratiques évaluatives associées à l'APC. Bien qu'ils soient interreliés, ces thèmes seront abordés indépendamment afin de permettre au lecteur de bien comprendre les différents concepts. Les pratiques évaluatives discutées dans cette section sont actuellement préconisées dans les programmes afin de préparer les futurs professionnels de la santé à agir avec compétence dans les différents milieux de soins. Pour chacun des thèmes, nous sensibiliserons le lecteur aux défis présents ou potentiels en ce qui concerne la qualité des pratiques évaluatives.

2.2.1 Évaluation des apprentissages

L'évaluation des apprentissages est un sujet complexe. Plusieurs concepts reliés à l'évaluation (p. ex. buts, enjeux, fonction) sont souvent confondus entre eux. Nous distinguerons ces termes afin de mieux comprendre les défis qui y sont rattachés. Tout d'abord, nous définirons le but de l'évaluation (formative et sommative), qui correspond à une manière plus traditionnelle de désigner le rôle de l'évaluation. Puis, nous décrirons les enjeux (faibles ou élevés) de l'évaluation qui représentent une vision plus récente de l'évaluation des apprentissages. Finalement, nous différencierons l'évaluation de l'apprentissage et l'évaluation pour l'apprentissage.

Buts de l'évaluation

L'évaluation peut être utilisée à des fins formatives ou sommatives (Bouldin & Wilkin, 2000; Schuwirth & van der Vleuten, 2019). L'évaluation formative est utilisée principalement pour communiquer une rétroaction à l'apprenant et l'informer sur sa performance (Peeters, 2017). L'évaluation formative s'effectue dans un contexte sécuritaire pour l'apprenant qui utilise la rétroaction pour progresser dans son apprentissage (Brown, 2019). Plus formelle, l'évaluation sommative vise à démontrer l'atteinte d'un niveau de compétence face à des critères précis (Brown, 2019). Habituellement, les résultats d'une évaluation sommative sont émis sous forme de nombre, de lettre ou de grade (Brown, 2019; Peeters, 2017). Que ce soit une évaluation formative ou sommative, les résultats obtenus mènent à des prises de décisions. Par exemple, l'évaluation formative peut amener l'enseignant à modifier son

approche pédagogique ou à offrir du soutien à un apprenant en difficulté (Peeters, 2017). L'évaluation sommative peut mener à l'admission dans un programme ou à l'obtention d'un diplôme. Il est recommandé de rédiger les buts (formatif ou sommatif) et les objectifs de l'évaluation de manière claire et compréhensible pour les apprenants (Assessment Reform Group, 2002; Dannefer, 2013; Wiliam, 2011).

Enjeux de l'évaluation

L'évaluation des apprentissages peut être considérée comme une évaluation de faible enjeu (p. ex. une évaluation formative) ou une évaluation à enjeu élevé (p. ex. certification ou obtention d'un diplôme) (Kogan, Hatala, Hauer, & Holmboe, 2017). Bien qu'il y ait des ressemblances entre le but de l'évaluation (formatif ou sommatif) et l'enjeu de l'évaluation (faible ou élevé), il ne s'agit pas de synonymes. Qu'elle soit formative ou sommative, l'évaluation peut être d'enjeu faible ou élevé. Les enjeux d'une évaluation sont déterminés par les actions découlant de l'interprétation des résultats. Par exemple, une évaluation sommative peut être de faible enjeu en raison d'une petite pondération à l'intérieur d'un cours ou d'un programme. Une évaluation formative, quant à elle, peut être considérée à enjeu élevé si elle engendre des répercussions importantes (p. ex. modifications dans le choix des méthodes pédagogiques ou rédaction d'un plan de remédiation pour un apprenant en difficulté).

Les conséquences découlant de l'interprétation des résultats d'une évaluation à enjeu élevé revêtent une plus grande importance pour l'apprenant et pour la société. Par exemple, dans le contexte de la pédagogie des sciences de la santé, la réussite d'un examen de certification permet à l'apprenant de donner des soins de manière indépendante. Ainsi, la qualité de l'évaluation des apprentissages est essentielle pour assurer la sécurité des patients (Cook & Lineberry, 2016; Lockyer et al., 2017).

Dans une évaluation programmatique (voir section 2.2.3), la performance de l'apprenant est documentée à plusieurs moments, à l'aide de différentes méthodes d'évaluation (Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2012; van der Vleuten, Schuwirth, Driessen, Govaerts, & Heeneman, 2015). Les expressions « évaluation formative » et « évaluation

sommative » sont utilisées moins fréquemment. L'évaluation de faible enjeu et à enjeu élevé se positionne plutôt sur un continuum (van der Vleuten & Heeneman, 2016; van der Vleuten et al., 2015). De multiples évaluations de faible enjeu génèrent des « photos » qui documentent la performance de l'apprenant (Schut et al., 2018; van der Vleuten & Heeneman, 2016). La rétroaction découlant de ces évaluations offre des opportunités d'apprentissages (Pugh & Regehr, 2016). Les évaluations de faible enjeu sont ensuite regroupées pour obtenir un portrait complet de l'apprenant. Puis, la ou les personnes responsables de l'évaluation s'appuieront sur l'ensemble des données pour prendre une décision impliquant un enjeu élevé (Bok et al., 2013; van der Vleuten & Heeneman, 2016; van der Vleuten et al., 2015; Watling & Ginsburg, 2019) comme c'est le cas à la fin d'une étape dans un programme de formation ou pour l'obtention d'un diplôme (Baird et al., 2017; Pugh & Regehr, 2016).

Évaluation pour l'apprentissage

L'évolution des pratiques évaluatives a mené à des changements dans leur intention (Wiliam, 2011). L'évaluation de l'apprentissage (*assessment of learning*) a pour objectif de vérifier l'efficacité de l'enseignement et l'atteinte d'un certain niveau d'apprentissage ou de compétence (Harrison & Wass, 2016; Scallon, 2004; Schuwirth & van der Vleuten, 2011; Wiliam, 2011). L'évaluation de l'apprentissage est critiquée pour certaines conséquences négatives chez l'apprenant (p. ex. l'apprentissage « par cœur » dans le but de passer un examen). Souvent, dans ce contexte, la rétroaction est non spécifique et les apprenants démontrent peu d'intérêt à améliorer les éléments plus problématiques une fois l'examen réussi (Pugh & Regehr, 2016). De plus en plus, l'évaluation de l'apprentissage est remplacée par l'évaluation pour l'apprentissage (*assessment for learning*).

L'évaluation pour l'apprentissage favorise l'acquisition d'apprentissages chez l'apprenant (Martinez & Lipson, 1989; Peeters, 2017; Pugh & Regehr, 2016; Schuwirth & van der Vleuten, 2011; Wiliam, 2011) et permet aux personnes responsables de l'évaluation de juger de la progression des apprentissages (Lockyer et al., 2017; van der Vleuten & Dannefer, 2012). Toutefois, comme il s'agit d'un changement récent dans le rôle de l'évaluation, la culture de l'évaluation pour l'apprentissage n'est pas encore intégrée chez les apprenants en

sciences de la santé (Tekian, Watling, Roberts, Steinert, & Norcini, 2017). Dans les programmes de médecine, par exemple, les apprenants sont habitués à être compétitifs et visent la réussite avec un « A » (Dannefer, 2013; Pugh & Regehr, 2016). Ces personnes ne perçoivent pas nécessairement l'évaluation comme une opportunité d'apprentissage. Dans une approche d'évaluation pour l'apprentissage, toutes les évaluations (formatives et sommatives) ont le potentiel de favoriser la progression de l'apprenant (Baird et al., 2017; Peeters, 2017; Pugh & Regehr, 2016). L'évaluation pour l'apprentissage ne réfère pas qu'à l'évaluation formative (Baird et al., 2017). L'évaluation sommative informe l'apprenant sur sa performance et lui indique comment s'améliorer (Peeters, 2017; Wiliam, 2011). Toujours en gardant en tête l'apprenant et ses apprentissages, l'évaluation est également utile pour renseigner les enseignants sur l'efficacité des méthodes pédagogiques utilisées (Wiliam, 2011).

L'apprentissage en contexte d'évaluation peut survenir à trois moments : 1) avant l'évaluation, 2) pendant l'évaluation et 3) après l'évaluation. Avant l'évaluation, l'apprenant peut être motivé par l'imminence de l'évaluation (Cilliers, Schuwirth, Herman, Adendorff, & van der Vleuten, 2012; Cilliers, Schuwirth, & van der Vleuten, 2012; Pugh & Regehr, 2016). L'apprentissage est également influencé par le contenu ou par le format de l'évaluation (Cilliers, Schuwirth, Herman, et al., 2012). La préparation de l'apprenant sera différente s'il s'agit d'un examen clinique objectif structuré (ECOS) ou d'un QCM (Pugh & Regehr, 2016). Pendant l'évaluation, l'apprentissage est possible, car l'action de tester oblige l'apprenant à retrouver les informations nécessaires pour répondre aux questions de l'examen. Plus l'apprenant est questionné régulièrement, plus il se construira un système de recherche d'informations efficace (Carrier & Pashler, 1992; Karpicke & Grimaldi, 2012; Larsen, Butler, & Roediger, 2013; Pugh & Regehr, 2016). Après l'évaluation, l'apprenant reçoit de la rétroaction qui l'informe sur ses forces et ses points à améliorer. La rétroaction doit toutefois être donnée correctement et bonifiée par des commentaires narratifs (Tekian et al., 2017; Wood, 2019).

De nombreux auteurs reconnaissent le lien étroit entre l'évaluation et l'apprentissage des apprenants (Berendonk et al., 2013; Cilliers, Schuwirth, Herman, et al., 2012; Dannefer &

Henson, 2007; Darling-Hammond & Snyder, 2000; DeLuca, 2011; Dijkstra et al., 2010; Hays, 2008; Johnson, 2007; Maclellan, 2004; Palm, 2008; Prideaux et al., 2011; Schuwirth & van der Vleuten, 2011; Shumway & Harden, 2003; van der Vleuten et al., 2005; van der Vleuten et al., 2010). C'est pourquoi chacune des décisions concernant les méthodes d'évaluation et le moment auquel elles se déroulent doivent être réfléchies et planifiées (Assessment Reform Group, 2002; Brown, 2019; Wiliam, 2011). Les personnes qui élaborent le programme de formation doivent intégrer des opportunités de rétroaction afin de maximiser l'apprentissage (Brown, 2019; Wood, 2019).

2.2.2 Rétroaction

La rétroaction est essentielle pour la progression des apprentissages de l'apprenant (Brown, 2019; Cooper, 2000; Coulby, Hennessey, Davies, & Fuller, 2011; Crossley, Humphris, & Jolly, 2002; Hays, 2008; Pugh & Regehr, 2016; Richardson, Resick, Leonardo, & Pearsall, 2009; Ulfvarson & Oxelmark, 2012; van der Vleuten et al., 2010; Watling & Ginsburg, 2019; Wiggins, 2011). La rétroaction est un moyen de minimiser les conséquences de l'évaluation sur l'apprenant en lui permettant de se réajuster, d'être guidé dans la bonne direction et de se préparer pour les prochaines évaluations. Pour les mêmes raisons, cette rétroaction doit être de qualité afin d'éviter des répercussions négatives pour l'apprenant.

La rétroaction peut être donnée sous forme numérique (p. ex. note ou lettre) ou narrative (p. ex. commentaires) (Tekian et al., 2017; van der Vleuten & Heeneman, 2016). La rétroaction doit être de qualité (Curran et al., 2012; Dannefer & Henson, 2007; Darling-Hammond & Snyder, 2000; Eva et al., 2015; Kogan, Holmboe, & Hauer, 2009; Rentschler, Eaton, Cappiello, McNally, & McWilliam, 2007; Rosen et al., 2008; van der Vleuten et al., 2010), ce qui signifie qu'elle doit être axée sur l'apprentissage, liée à la tâche demandée ainsi qu'aux critères de correction et être compréhensible pour l'apprenant (p. ex. l'utilisation d'un vocabulaire accessible à l'apprenant) (Gibbs, 2019). La rétroaction est transmise à l'apprenant afin qu'il comprenne bien le but de l'évaluation ainsi que l'écart entre la performance attendue et sa performance actuelle (Assessment Reform Group, 2002; Lockyer et al., 2017; Tekian et al., 2017; Wiliam, 2011; Wood, 2019). L'apprenant doit également comprendre comment il pourrait mettre en place des stratégies pour combler l'écart et

améliorer sa performance (Watling & Ginsburg, 2019; Wiliam, 2011; Wood, 2019). Les enseignants et le programme de formation doivent offrir des opportunités de rétroaction fréquentes (Eva et al., 2015; Gibbs, 2019) et assez rapidement après l'évaluation pour être utile (Gibbs, 2019; Isaacson & Stacy, 2009; Johnson, 2007; Rentschler et al., 2007; Rosen et al., 2008).

Défis reliés à la rétroaction de qualité

Bien que les avantages de la rétroaction soient démontrés, certains défis y sont associés. La fréquence optimale pour recevoir de la rétroaction, le délai avant d'obtenir une rétroaction ou encore la composition exacte de la rétroaction demeurent peu documentés à ce jour dans le domaine des sciences de la santé. Il peut être difficile d'atteindre l'équilibre entre la quantité et la fréquence de rétroaction et sa qualité (Gibbs, 2019). Par exemple, des réponses automatisées peuvent être obtenues immédiatement après un QCM complété par ordinateur. Toutefois, il se peut que la rétroaction ne soit pas aussi complète que si elle était donnée par une personne évaluatrice. Aussi, la sécurité du contenu de l'examen demeure un enjeu relié à la rétroaction. Dans certains cas, comme pour les ECOS, il peut être risqué de donner de la rétroaction sans divulguer le contenu de l'examen, rendant impossible son utilisation pour une prochaine cohorte (Pugh & Regehr, 2016). Les coûts associés à la rétroaction, principalement en raison des ressources humaines exigées (p.ex. temps exigé pour donner de la rétroaction à l'apprenant et pour documenter la progression), peuvent également représenter un enjeu pour les établissements d'enseignement (van der Vleuten & Heeneman, 2016).

La rétroaction de type narratif semble être plus efficace pour favoriser l'apprentissage des apprenants (Driessen & Scheele, 2013; Tekian et al., 2017). Toutefois, pour les évaluateurs, cela peut représenter une source de stress (Harrison & Wass, 2016; Pugh & Regehr, 2016) pour deux raisons principales. Premièrement, une rétroaction de mauvaise qualité peut engendrer des conséquences sur l'apprenant, c'est-à-dire que les commentaires peuvent avoir des répercussions sur la confiance en soi des apprenants (Assessment Reform Group, 2002; Pugh & Regehr, 2016). Deuxièmement, une rétroaction de qualité exige des efforts et du temps de la part du personnel enseignant (Hawkins et al., 2015). Aussi, il n'est pas toujours

possible pour les personnes responsables de l'évaluation d'être disponibles pour donner une rétroaction dans les milieux cliniques (Wood, 2019). Il est d'ailleurs recommandé de former et de soutenir les enseignants afin qu'ils puissent transmettre de la rétroaction de qualité aux apprenants (Assessment Reform Group, 2002; Tekian et al., 2017; van der Vleuten & Heeneman, 2016; van der Vleuten et al., 2010).

2.2.3 *Évaluation programmatique*

Dans une approche où l'évaluation favorise l'apprentissage, l'apprenant est au cœur de toutes les réflexions et les décisions concernant l'évaluation des apprentissages, notamment les méthodes d'évaluation, les moments d'évaluation, la rétroaction et l'observation de la performance par plusieurs évaluateurs. L'évaluation programmatique (*programmatic assessment*), aussi appelée programme d'évaluation, est constituée de différentes sources d'informations permettant de collecter des « photos » de la performance d'un apprenant à différents moments (Dijkstra et al., 2010; Pugh & Regehr, 2016; Schuwirth & van der Vleuten, 2011; Schuwirth & van der Vleuten, 2012; Timmerman & Dijkstra, 2017; van der Vleuten & Dannefer, 2012; van der Vleuten et al., 2012). L'utilisation de plusieurs sources de données sur l'apprenant rend possible la triangulation des données (Baartman et al., 2013; Schuwirth & van der Vleuten, 2012; Shumway & Harden, 2003; van der Vleuten et al., 2005; van der Vleuten et al., 2010). L'interprétation de toutes les « photos » permet d'obtenir un portrait plus juste de la performance de l'apprenant que les résultats d'une seule évaluation (Baartman et al., 2013; Dijkstra et al., 2010; Pugh & Regehr, 2016; Schuwirth & van der Vleuten, 2012; Timmerman & Dijkstra, 2017; van der Vleuten et al., 2005). Les décisions découlant de l'interprétation des résultats de l'évaluation seront plus défendables aux yeux de l'apprenant, de l'établissement ou des instances professionnelles.

L'analogie du diagnostic médical utilisée par Schuwirth et van der Vleuten (2012; 2019) permet de bien comprendre l'évaluation programmatique. Avant de poser un diagnostic (p. ex. hypertension), le médecin collecte des données à plusieurs occasions et avec différentes stratégies (p. ex. histoire du patient, résultats de laboratoire, examen physique et signes vitaux). La personne évaluatrice fera de même pour obtenir un portrait plus complet de la performance de l'apprenant.

Choix des méthodes d'évaluation

Les méthodes d'évaluation doivent être sélectionnées judicieusement (Prideaux et al., 2011; Roberts et al., 2006; van der Vleuten et al., 2005; van der Vleuten et al., 2010). Un seul format d'évaluation (p. ex. un examen écrit) n'est pas suffisant pour mesurer le niveau de compétence de l'apprenant (Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2005). Le programme d'évaluation est construit de manière à ce que les faiblesses de certaines méthodes d'évaluation soient compensées par les forces des autres (Brydges, Ginsburg, & Hatala, 2015; Dijkstra et al., 2010; Hawkins et al., 2015; Norcini & Zaidi, 2019; Schuwirth & van der Vleuten, 2012, 2019; Timmerman & Dijkstra, 2017). La combinaison de méthodes qui engendrent des données quantitatives (p. ex. QCM) et qualitatives (p. ex. portfolio) enrichissent les informations sur la performance de l'apprenant (Lockyer et al., 2017). Le choix des méthodes intégrées dans le programme doit correspondre aux ressources humaines disponibles dans le milieu d'enseignement comme la disponibilité, le nombre ou l'expérience des évaluateurs (Hawkins et al., 2015; Lockyer et al., 2017; van der Vleuten et al., 2005). Les ressources financières sont également à considérer et peuvent influencer les décisions. Par exemple, il peut s'avérer coûteux et exigeant de développer plusieurs ECOS (Schuwirth & van der Vleuten, 2019).

Choix des moments d'évaluation

Le programme d'évaluation comporte plusieurs moments d'évaluation (Albanese et al., 2010; Baartman et al., 2013; Callwood, Cooke, & Allan, 2014; Cook et al., 2015; Dannefer, 2013; Darling-Hammond & Snyder, 2000; Eva et al., 2015; Kasanda, Mitonga, Vei, & Zimba, 2013; Neary, 2001; Schuwirth & van der Vleuten, 2011; Shumway & Harden, 2003; van der Vleuten et al., 2005; Vriend Van Duinen, 2005). Il doit intégrer des évaluations ayant une progression dans leur complexité et offrir des opportunités fréquentes d'apprentissage (p. ex. rétroaction) (Brown, 2019; Epstein, 2007; Norcini et al., 2011; van der Vleuten, Schuwirth, Driessen, Govaerts, & Heeneman, 2015; Wood, 2019). Toutefois, les personnes responsables de l'évaluation doivent tenir compte de la fatigue de l'apprenant pour choisir la fréquence, le nombre et le moment des évaluations (Hodges, 2013; Lockyer et al., 2017; Norcini et al., 2011). Lorsque les évaluations formatives sont nombreuses, l'apprenant peut

tout de même ressentir un stress comme s'il s'agissait à chaque fois d'une évaluation sommative (Harrison & Wass, 2016).

Multiples observateurs

L'observation ou la supervision de la performance de l'apprenant par plus d'une personne enrichissent la rétroaction offerte à l'apprenant (Fitzgerald et al., 2016) et consolident les décisions découlant de l'interprétation des résultats de l'évaluation (Baartman et al., 2007; Fitzgerald et al., 2016; Johnson, 2007; Lockyer et al., 2017; Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2005; Wiggins, 2011). Les données recueillies par plusieurs personnes équilibrent la clémence et la sévérité susceptibles d'être présentes chez les personnes évaluatrices (Lockyer et al., 2017).

Défis de l'évaluation programmatique

L'utilisation d'un programme d'évaluation soulève toutefois plusieurs questions. Comment peut-on interpréter l'abondance des informations générées par les multiples évaluations (Hawkins et al., 2015)? Comment peut-on interpréter la globalité des données recueillies à l'intérieur du programme d'évaluation (Hawkins et al., 2015; Schuwirth & Ash, 2013)? Est-il possible de combiner des résultats numériques aux résultats narratifs? Combien de données sont suffisantes pour prendre une décision sur la compétence d'un apprenant (Norman, Norcini, & Bordage, 2014)? Que faire si les données recueillies sont divergentes? La plupart de ces questions demeurent malheureusement sans réponse définitive et sont en partie responsables des préoccupations concernant les limites des processus de validation existants (voir section 2.4 sur les processus de validation). Ces préoccupations ont mené les chercheurs à considérer la validité autrement, c'est-à-dire avec une vision plus holistique et la volonté d'interpréter l'ensemble des résultats obtenus sur la performance d'un apprenant dans une évaluation programmatique.

2.2.4 Approche par compétences

Le programme d'évaluation est élaboré dans une perspective d'approche par compétences (APC), c'est-à-dire que le programme permet de documenter la progression des compétences de l'apprenant. Dans une APC, l'apprenant doit démontrer qu'il possède les compétences

nécessaires pour exercer sa profession. Ces compétences sont des éléments explicites et décontextualisés telles que la connaissance d'un sujet ou la capacité à poser un geste (Govaerts & van der Vleuten, 2013; Grant, 2019). Les personnes qui développent et administrent les programmes de formation ont la responsabilité sociale de s'assurer que les compétences enseignées aux futurs professionnels sont en cohérence avec les besoins évolutifs de la société (p. ex. vieillissement de la population, complexité des soins, accès aux soins de santé) (Boelen, 2011, 2018; Boyd et al., 2018; Frank, Snell, Englander, & Holmboe, 2017; Hodges & Lingard, 2012). Le public (p. ex. les patients, les familles) doit avoir confiance dans la qualité des programmes qui forment et évaluent les futurs professionnels de la santé (Norman et al., 2014; Surjadi, Stringari-Murray, & Saxe, 2019).

Dans une APC, il est important de documenter le développement des apprenants et d'observer la performance de ces derniers. Par conséquent, l'APC a nécessité le développement de nouvelles pratiques évaluatives (p. ex. test de progression) et une plus grande utilisation de méthodes d'évaluation générant des données qualitatives (p. ex. portfolio, évaluation en milieu de travail). Dans cette section, nous présenterons les principaux changements dans les pratiques évaluatives qui s'inscrivent bien dans une APC avec leurs avantages, mais également leurs limites. Ces limites sont documentées dans les écrits scientifiques et ont contribué à l'émergence de la conceptualisation de la validité en tant qu'impératif social. Tout d'abord, nous décrirons ce qui caractérise les jalons de compétences et les activités professionnelles fiables. Ensuite, nous aborderons l'évaluation de la progression de l'apprenant en détaillant deux méthodes d'évaluation intégrées dans plusieurs programmes de formation en sciences de la santé soit, le test de progression et le portfolio. Finalement, nous aborderons l'observation et l'évaluation de la performance de l'apprenant par la description de deux méthodes évaluatives fréquemment utilisées, c'est-à-dire l'ECOS et l'évaluation en milieu de travail.

Jalons de compétences et activités professionnelles fiables

Dans les programmes de formation, les compétences sont souvent réduites en sous-éléments, alors qu'elles devraient être considérées dans un ensemble (Hodges, 2013; Holmboe, Sherbino, Englander, Snell, & Frank, 2017; Lockyer et al., 2017; Scallion, 2004; van der

Vleuten et al., 2005). Afin d'avoir une meilleure vision intégrative des compétences de l'apprenant et de sa progression, les programmes de médecine ont instauré les jalons (*milestones*) et les activités professionnelles confiées (*entrustable professional activity*) (Lockyer et al., 2017; Royal College of Physicians and Surgeons of Canada, 2019b; Surjadi et al., 2019; Ten Cate, 2013).

Les jalons de compétences sont des étapes dans le développement de compétences (Ten Cate, 2013; Wagner & Reeves, 2015). Ces jalons guident la personne qui accompagne l'apprenant sur le niveau de supervision à fournir pendant un geste ou une action (c.-à-d. observation par l'apprenant sans exécution du geste, l'exécution d'un geste avec supervision directe ou supervision à distance, l'exécution d'un geste sans supervision) (Norcini & Zaidi, 2019; Schultz, Griffiths, & Lacasse, 2015; Ten Cate, 2013). La progression dans le degré de supervision nécessite plusieurs observations (Peters, Holzhausen, Boscardin, ten Cate, & Chen, 2017; Ten Cate, 2013). Les décisions de supervision sont influencées par 1) l'apprenant (p. ex. son niveau d'expertise), 2) le superviseur (p. ex. plus ou moins strict), 3) le contexte (p. ex. les ressources disponibles) et 4) la nature du geste (p. ex. complexité) (Ten Cate, 2013).

Les activités professionnelles confiées sont utilisées pour décrire les compétences dans un contexte clinique (Ten Cate, 2013; Wagner & Reeves, 2015). Il s'agit d'une liste de tâches pouvant être confiées à l'apprenant (Wagner & Reeves, 2015) comme, par exemple, procéder à l'examen physique de l'abdomen ou encore administrer un médicament intraveineux. Une fois que l'apprenant acquiert un niveau de compétence, l'évaluateur lui accorde sa confiance pour poser un geste sous supervision ou sans supervision (Sklar, 2015). L'apprenant doit également démontrer sa capacité de demander de l'aide au besoin (Surjadi et al., 2019; Ten Cate, 2013).

Élaborés en premier lieu pour les programmes de résidence, les jalons et les activités professionnelles confiées sont maintenant pratique courante dans les programmes prégradués de médecine (Peters, Holzhausen, Boscardin, ten Cate, & Chen 2017). La possibilité d'inclure les jalons et les activités professionnelles confiées est également

examinée par les programmes de formation comme ceux développés pour les futures infirmières praticiennes (Surjadi et al., 2019) ou en pédagogie des sciences de la santé (Fitzgerald et al., 2016). Sans être appelée « jalons » ou « activités professionnelles fiables », la progression des apprentissages représente une caractéristique de nombreux programmes de formation (Albanese et al., 2010; Barton, Armstrong, Preheim, Gelmon, & Andrus, 2009; Croft, Gilligan, Rasiah, Levett-jones, & Schneider, 2007; Scallon, 2004; Stevens, Hyde, Knight, Shires, & Alexander, 2017; Tardif, 2006).

Inquiétudes liées aux jalons de compétences et aux activités professionnelles fiables

Comme les activités professionnelles fiables sont élaborées sous forme de listes, il est possible que l'évaluateur vérifie simplement si l'apprenant répond aux critères (Eva et al., 2015; Peters et al., 2017), ce qui diminue les opportunités de rétroaction. Il se peut également que les apprenants ciblent la compétence minimale sans viser l'excellence (Peters et al., 2017).

Une critique des jalons de compétences et des activités professionnelles fiables concerne la progression variable dans le temps, d'un individu à l'autre. En d'autres mots, un apprenant peut compléter un jalon avant un autre apprenant de même année scolaire (Fitzgerald et al., 2016; Hawkins et al., 2015; Norman et al., 2014). Cette variabilité entre les apprenants complique la tâche du personnel enseignant dans les programmes de formation ou encore celle des instances professionnelles (Hawkins et al., 2015; Holmboe et al., 2017) concernant, par exemple, la fin du programme pour un apprenant ou le moment de sa certification.

De par les pratiques évaluatives associées (observation directe, évaluations nombreuses, rétroaction fréquente), les jalons de compétences et les activités professionnelles fiables exigent beaucoup de temps et de ressources (Hawkins et al., 2015). De plus, si la description des compétences, des jalons et des tâches exigés chez l'apprenant manquent de clarté, de la confusion peut survenir chez toutes les parties prenantes (Hawkins et al., 2015; Norman et al., 2014). Aussi, les qualités psychométriques des méthodes utilisées pour évaluer la

progression de l'apprentissage des apprenants (p. ex. portfolios) provoquent des questionnements chez les utilisateurs et les chercheurs.

L'évaluation de la progression de l'apprenant

La progression de l'apprenant peut être évaluée par différentes méthodes d'évaluation. Nous avons choisi de présenter deux d'entre elles en raison de leur popularité et de leur plus récente intégration dans les programmes de formation en sciences de la santé soit, le test de progression et le portfolio.

Administré régulièrement, le test de progression (*progress test*) est un examen qui porte sur l'ensemble du contenu d'un programme (Blake et al., 1996; Pugh & Regehr, 2016; van der Vleuten & Heeneman, 2016; Wrigley, van der Vleuten, Freeman, & Muijtjens, 2012). Il peut s'agir, par exemple, d'un examen administré chaque année à l'ensemble des résidents en médecine interne, permettant d'évaluer leur progression (Pugh & Regehr, 2016). Le test de progression est considéré comme une évaluation de faible enjeu puisque son rôle est de soutenir l'apprenant dans ses apprentissages, de fournir une rétroaction et de favoriser la préparation pour l'examen de certification (Harrison & Wass, 2016; Pugh & Regehr, 2016). Comme le contenu évalué est diversifié et volumineux, il est plus difficile pour l'apprenant de mémoriser « par cœur » les informations, favorisant ainsi l'apprentissage en profondeur (Harrison & Wass, 2016; Pugh & Regehr, 2016). Les défis liés à l'utilisation du test de progression sont, notamment, le format écrit qui mesure principalement l'acquisition des connaissances de l'apprenant (Pugh, Touchie, Wood, & Humphrey-Murto, 2014) et la vaste étendue du contenu à étudier qui pourrait décourager les apprenants (Pugh & Regehr, 2016).

Le portfolio est une autre méthode d'évaluation pouvant être utilisée pour mesurer la progression de l'apprenant (Driessen, 2017; Eva et al., 2015; Peters et al., 2017; van der Vleuten et al., 2012). Le portfolio est utile, car il permet de regrouper au même endroit les informations sur les compétences de l'apprenant. (Dannefer & Henson, 2007; Norcini & Zaidi, 2019; Schuwirth & van der Vleuten, 2019; van der Vleuten et al., 2015; van der Vleuten et al., 2010; Van Tartwijk & Driessen, 2009). Dans les programmes de formation en sciences de la santé, le portfolio est habituellement utilisé pour soutenir l'apprenant dans ses

apprentissages et pour évaluer son développement (Driessen, 2017; Peeters, 2017). Bien que contesté en raison de sa subjectivité, le portfolio regroupe des données qui enrichissent le portrait de la performance de l'apprenant (Lockyer et al., 2017).

Le temps exigé pour compléter le portfolio peut sembler important pour l'apprenant (Driessen, 2017). Certains portfolios sont volumineux et l'évaluateur peut passer beaucoup plus de temps à lire qu'à donner de la rétroaction utile à l'apprenant (Driessen, 2017; Van Tartwijk & Driessen, 2009). Le but et le format du portfolio varient énormément (format électronique ou non, quantité d'informations recueillies) (Norcini & Zaidi, 2019; Peeters, 2017; Scallon, 2004; Van Tartwijk & Driessen, 2009). De plus, certaines formes de portfolio comportent une réflexion de l'apprenant, ce qui rend difficile la notation du contenu (Driessen, 2017; Hodges, 2015; Schuwirth & van der Vleuten, 2019; van der Vleuten et al., 2015; Van Tartwijk & Driessen, 2009). Les analyses psychométriques traditionnelles pour en vérifier le degré de validité et de fidélité s'avèrent limitées (Baartman et al., 2013; Scallon, 2004; Van Tartwijk & Driessen, 2009). Des auteurs sont d'avis que l'utilisation des critères de rigueur de recherche qualitative (p. ex. crédibilité et la fiabilité) pourrait être une manière de s'assurer de la qualité du portfolio (Van Tartwijk & Driessen, 2009).

Observation et évaluation de la performance

Dans une APC, il est de mise de favoriser l'utilisation de méthodes d'évaluation en situation authentique, afin que les outils mesurent la performance de l'apprenant dans un contexte se rapprochant de la réalité clinique (Darling-Hammond & Snyder, 2000; Neary, 2001; Palm, 2008; Prigent, Bernard, & Kozanitis, 2009; Scallon, 2004; Scoles et al., 2003; Shumway & Harden, 2003; Tardif, 2006; van der Vleuten et al., 2010; van der Vleuten & Schuwirth, 2005; Wiggins, 2011). L'observation et l'évaluation de la performance de l'apprenant a exigé l'élaboration de nouvelles méthodes d'évaluation. Les principales méthodes permettant l'observation et l'évaluation de la performance de l'apprenant que l'on retrouve dans les programmes de formation en sciences de la santé sont l'ECOS et l'évaluation en milieu de travail.

Bien que ce ne soit pas une nouvelle méthode d'évaluation, l'ECOS est utilisé fréquemment dans les APC. L'ECOS rend possible l'évaluation de la performance d'un apprenant en contrôlant certains aspects comme une observation plus structurée dans le contexte de situations cliniques prédéterminées (Schuwirth & van der Vleuten, 2019). Une station d'ECOS permet à tous les apprenants d'intervenir auprès d'un même patient standardisé ou d'un mannequin à l'intérieur d'une période de temps prédéterminée (Schuwirth & van der Vleuten, 2019). Cette méthode d'évaluation permet à l'apprenant de démontrer plusieurs compétences simultanément (p. ex. la communication avec le patient pendant l'auscultation pulmonaire). Les évaluateurs sont également standardisés et ils utilisent une grille pour noter la performance de l'apprenant (Schuwirth & van der Vleuten, 2019). Toutefois, l'ECOS nécessite l'implication de plusieurs personnes et requiert une logistique complexe, ce qui en fait une méthode coûteuse (Schuwirth & van der Vleuten, 2019).

L'évaluation en milieu de travail (*workplace-based assessment*) est un exemple d'évaluative authentique, car c'est ce qui se rapproche le plus de la réalité des futurs professionnels de la santé (Crossley & Jolly, 2012; Norcini & Zaidi, 2019). L'évaluation en milieu de travail repose sur l'observation de la performance d'un apprenant par un évaluateur à l'aide d'un outil (Harris et al., 2017). Différents outils servent à évaluer la performance d'apprenants dans le milieu de travail. Par exemple, dans les programmes de médecine, l'outil mini-CEX a été développé pour documenter la performance d'un apprenant lors d'une observation directe pendant l'entrevue avec le patient (Schuwirth & van der Vleuten, 2019). L'évaluation multisource (aussi appelée évaluation à 360 degrés) permet de documenter la performance de l'apprenant, de manière indirecte, selon le jugement de plusieurs acteurs tels que l'évaluateur, les collègues, les patients ou encore les autres professionnels de la santé (Crossley & Jolly, 2012; Schuwirth & van der Vleuten, 2019).

Plusieurs enjeux liés à l'observation et à l'évaluation de la performance sont mentionnés dans les écrits scientifiques. Premièrement, certaines compétences ne peuvent pas être décrites facilement par des caractéristiques spécifiques ou mesurables (p. ex. le professionnalisme ou le travail d'équipe) (Albanese et al., 2010; Hawkins et al., 2015). Deuxièmement, la diversité des milieux et les situations cliniques rencontrées complexifient la mesure de la qualité d'une

méthode d'évaluation en milieu de travail (Govaerts, Schuwirth, van der Vleuten & Muijtens, 2011; Govaerts & van der Vleuten, 2013). Troisièmement, les outils utilisés pour évaluer l'apprenant en milieu de travail incluent souvent un contenu narratif (p. ex. évaluation de stage). Ces informations alimentent l'évaluateur lors de la rétroaction et lors de la prise de décisions (Hawkins et al., 2015). Toutefois, les analyses psychométriques existantes s'avèrent insuffisantes pour en assurer la qualité (Hawkins et al., 2015; Kogan et al., 2009).

Quatrièmement, les évaluations en situation authentique nécessitent l'observation et l'interprétation des données par une personne évaluatrice (Baartman et al., 2013; Govaerts & van der Vleuten, 2013; Grant, 2019). Bien que certaines grilles ou rubriques d'évaluation aient fait l'objet de processus de validation (Govaerts et al., 2011), la subjectivité ne pourra jamais être complètement éliminée de l'évaluation (Govaerts et al., 2011; van der Vleuten et al., 2010). En effet, il a été démontré que les évaluateurs ne sont pas toujours d'accord lorsqu'ils évaluent la même tâche chez le même apprenant (Gingerich, Ramlo, van der Vleuten, Eva, & Regehr, 2017; Norcini & Zaidi, 2019; St-Onge, Chamberland, Lévesque, & Varpio, 2016). D'un point de vue psychométrique, la variabilité entre les évaluateurs est vue comme une erreur de mesure (Gingerich, van der Vleuten, Eva, & Regehr, 2014). D'un autre point de vue, si plusieurs évaluateurs n'ont pas la même interprétation de la performance d'un apprenant, il se peut que cette différence enrichisse le portrait de la performance de l'apprenant (Gingerich, Kogan, Yeates, Govaerts, & Holmboe, 2014; Gingerich, van der Vleuten, et al., 2014; Hodges, 2013; Lockyer et al., 2017).

En résumé, l'évaluation des apprentissages est au cœur de la formation des futurs professionnels de la santé. Dans une APC, chacune des décisions concernant l'évaluation des apprentissages (p.ex. méthodes d'évaluation, rétroaction, fréquence des observations) doit être réfléchi afin de permettre à l'apprenant de progresser dans son développement. Toutefois, les pratiques évaluatives utilisées soulèvent des questionnements, notamment à propos de la justesse de l'interprétation des résultats ou de la faisabilité de leur mise en œuvre. Certaines pratiques évaluatives (p.ex. l'évaluation programmatique et l'évaluation longitudinale) sont récentes et utilisées depuis peu de temps dans différents établissements d'enseignement. Par conséquent, l'avancement des connaissances sur le sujet est à ses débuts.

2.3 Processus de validation

Dans les programmes en sciences de la santé, les pratiques évaluatives sont utilisées pour poser un jugement sur la performance d'un apprenant. Pour limiter les conséquences négatives sur l'apprenant et sur la société, les personnes qui sont chargées de l'évaluation des apprentissages ont la responsabilité de s'assurer de la qualité des pratiques évaluatives. Un processus de validation sert à en vérifier la qualité. Le processus de validation est un processus continu qui nécessite des preuves provenant de plusieurs sources (Streiner, Norman, & Cairney, 2015), soit par le recours à différentes stratégies de validation qui peuvent être quantitatives (c.-à-d. analyses psychométriques) et qualitatives. Ces stratégies servent à collecter les preuves de validité qui appuient l'interprétation des résultats et les décisions qui en découlent (Kane, 2006 ; Messick, 1989 ; Sireci & Sukin, 2013). Dans cette section de la recension des écrits, nous décrirons brièvement les différentes conceptualisations de la validité (classique et modernes) en détaillant davantage les théories modernes utilisées actuellement dans les programmes en pédagogie des sciences de la santé.

2.3.1 Conceptualisations de la validité

La conceptualisation classique de la validité, appelée également « trinité de la validité », provient, entre autres, des résultats des travaux de Cronbach et Meehl en 1995 qui décrivent la validité par trois éléments : critère, contenu et construit (Goodwin, 1997; Sireci & Sukin, 2013; Streiner et al., 2015). La validité de critère est utilisée pour mesurer la qualité d'un outil à l'aide d'analyses qui documentent la valeur prédictive et/ou la validité convergente d'un outil (Kane, 2006). La validité de contenu sert à examiner la représentativité du contenu au travers des items inclus dans l'outil (Taylor, 2013). La validité de construit réfère à la justesse avec laquelle un outil permet d'obtenir des résultats conformes à la théorie lui étant sous-jacente (DeVellis, 2016; Streiner et al., 2015). La première édition du livre de référence *Standards for Educational and Psychological Testing* détaillait cette vision classique de la validité (AERA, APA, & NCME, 1966).

Dans les années 1990, la conceptualisation classique de la validité a été remise en question. Les principales critiques étant que les trois éléments de la conceptualisation classique de la validité (contenu, construit et critères) correspondent tous à de la validité de construit

(Messick, 1989, 1995; Sireci & Sukin, 2013). Selon les théories modernes de validité, une méthode d'évaluation ne peut pas être valide en soi, ce sont plutôt les interprétations ou les décisions issues de ses résultats qui doivent être validées (Borsboom, Mellenbergh, & van Heerden, 2004; Chapelle, Enright, & Jamieson, 2010; Messick, 1995; Sireci & Sukin, 2013; Streiner et al., 2015). En 1995, Messick (1995) a décrit la théorie unifiée de la validité. En 2006, Kane (2006) a proposé une approche de validation fondée sur l'accumulation des preuves. Bien que Messick (1995) et Kane (2006) aient proposé deux théories distinctes, celles-ci présentent des similitudes. Pour les deux auteurs, un processus de validation nécessite la collecte de preuves de validité.

Encore plus récemment, Cizek (2012, 2016b) distingue l'utilisation d'un test et les inférences d'un test à travers sept dimensions (raisonnement, enquête, focus, tradition, mandats, temps, responsabilité). Par exemple, la responsabilité de la justification de l'utilisation d'un test incombe aux utilisateurs du test et aux décideurs (p. ex. instances professionnelles) alors que les personnes qui élaborent le test sont responsables de la qualité des inférences. Par son argumentaire, Cizek (2012, 2016b) soutient qu'il est nécessaire de réaliser deux processus de validation séparément, c'est-à-dire récolter des preuves de validité pour justifier l'utilisation d'un test et d'autres preuves pour vérifier la qualité des inférences.

Nous résumons les caractéristiques des principales conceptualisations de la validité dans le tableau 2. Bien qu'il existe d'autres conceptualisations de la validité, nous avons sélectionné ces auteurs (Messick et Kane) puisqu'ils sont les plus connus dans le contexte de la pédagogie des sciences de la santé.

Tableau 2
Conceptualisations de la validité

Théorie	Conceptualisation classique de la validité	Théories modernes	
Appellation	Trinité de la validité (Holy Trinity)	Théorie unifiée de la validité (<i>unitary concept of construct validity framework</i>)	Validation fondée sur l'accumulation des preuves (<i>argument-based rationale</i>)
Auteurs principaux	AERA, APA, NCME (1966) ¹	Messick (1989,1995)	Kane (2006)
Éléments de la validité	Validité de contenu Validité de construit Validité de critère	Preuve de contenu Complexité cognitive Structure interne Généralisabilité Aspects externes Conséquences	Argument interprétatif : • Score • Généralisation • Extrapolation • Décision Argument de validité

2.3.2 Théorie unifiée de la validité

La conceptualisation du processus de validation selon Messick (1989, 1995) représente une vision unifiée, fondée sur l'accumulation de preuves et sur l'importance des inférences (*unitary concept of construct validity framework*). Messick (1989, 1995) précise que la validation est un processus continu nécessitant l'accumulation de plusieurs preuves pour obtenir une interprétation juste du résultat de l'évaluation. La théorie unifiée de Messick (1989, 1995) comprend six preuves de validité (tableau 3) qui sont 1) la preuve de contenu, 2) la complexité cognitive, 3) la structure interne, 4) la généralisabilité, 5) les aspects externes et 6) les conséquences.

En 1999, lors de la publication du livre de référence *Standards for Educational and Psychological Testing*, l'*American Educational Research Association* (AERA), l'*American Psychological Association* (APA) et le *National Council on Measurement in Education* (NCME) ont adapté la théorie de Messick afin de la rendre plus accessible aux utilisateurs. La généralisabilité et les aspects externes ont été regroupés sous une même preuve nommée « relation avec les autres variables ». Bien que cette adaptation composée de cinq preuves de validité soit la plus fréquemment utilisée dans les écrits, nous avons choisi de décrire les six preuves originales décrites par Messick (1989, 1995).

¹ *American Educational Research Association* (AERA)
American Psychological Association (APA)
National Council on Measurement in Education (NCME)

La *preuve de contenu* correspond à la pertinence, à la représentativité et à la clarté du contenu de l'évaluation (Messick, 1995). Cette preuve peut être mesurée par le jugement d'experts obtenu de manière qualitative (p. ex. entrevues, discussions de groupes focalisées) ou de manière quantitative (p. ex. indice de validité de contenu – IVC). La personne responsable de l'évaluation peut également utiliser une table de spécification pour avoir une vue d'ensemble de tout le contenu évalué (AERA, APA & NCME, 2014; Sireci & Sukin, 2013). La *complexité cognitive* est une preuve documentée pour démontrer le lien entre le processus cognitif de l'apprenant et sa performance (AERA et al., 2014; Sireci & Sukin, 2013). Cette preuve est difficile à mesurer, car elle ne peut être observée directement. Les stratégies possibles sont des entrevues ou des groupes de discussion focalisée avec les apprenants, l'analyse du processus cognitif de l'apprenant pendant l'évaluation par une réflexion à voix haute ou encore l'analyse du temps de réponse (AERA et al., 2014; Sireci & Sukin, 2013). La preuve de *structure interne* sert à mesurer les liens entre les items et le construit de l'évaluation. Les analyses psychométriques sont nécessaires pour mesurer la consistance interne, la dimensionnalité et l'invariance dans la mesure. La *généralisabilité* est une preuve de validité qui vise à vérifier si les résultats sont transférables auprès de différentes populations ou dans différents contextes. Par exemple, les personnes responsables d'un examen de certification effectuent une validation de la langue et du contexte culturel pour ainsi s'assurer qu'un examen élaboré pour une population francophone est aussi approprié pour une population anglophone (Streiner et al., 2015). Les *aspects externes* concernent les relations entre le résultat de l'évaluation et les autres variables. La personne responsable de l'évaluation peut récolter des données pour vérifier la validité convergente ou divergente par des analyses psychométriques. Par exemple, dans les programmes de médecine, un nouvel examen d'admission moins coûteux pourrait être comparé à un examen d'admission possédant un bon niveau de validation, mais nécessitant plus de ressources humaines et financières. Les *conséquences* sont documentées pour mesurer les répercussions de l'interprétation des résultats de l'évaluation sur l'apprenant. Les conséquences peuvent être positives ou négatives, prévues ou imprévues. Comme l'évaluation est toujours utilisée dans un but précis, le résultat d'un test déclenche souvent des actions (Messick, 1989), telles que l'admission dans un programme ou l'obtention d'un diplôme (Hubley & Zumbo, 2011). Les stratégies de validation concernant la preuve de conséquences suggérées dans *Standards for*

Educational and Psychological Testing (AERA, APA & NCME, 1999) incluent l'utilisation de méthodes quantitatives et qualitatives. Le chercheur peut, par exemple, documenter le taux de réussite des apprenants, mesurer l'anxiété liée à l'évaluation ou encore interroger les apprenants à propos des répercussions de l'évaluation sur leur motivation à s'investir dans leur étude. Pour Messick (1995) et AERA et al. (1999, 2014), il importe que les conséquences ne proviennent pas d'une sous-représentation du construit (*construct underrepresentation*) ou d'un construit non pertinent (*construct irrelevant variance*) (Hubley & Zumbo, 2011; Messick, 1995).

Tableau 3
Théorie unifiée de la validité

Type de preuves de validité	Description	Exemples de stratégies de validation
Contenu	Construction des items <ul style="list-style-type: none"> • Pertinence • Représentativité • Clarté 	Jugement d'expert Table de spécification
Complexité cognitive	Processus cognitif de l'apprenant pendant l'évaluation	Entrevues et groupes de discussion focalisée Raisonnement à voix haute (<i>think aloud protocol</i>) Analyse du temps de réponse
Structure interne	Liens entre les items et le construit de l'évaluation <ul style="list-style-type: none"> • Consistance interne • Dimensionnalité • Invariance dans la mesure 	Exemples d'analyses psychométriques <ul style="list-style-type: none"> • Théorie de réponse aux items • Analyses multivariées
Généralisabilité	Transférabilité de l'évaluation auprès de différentes populations ou dans différents contextes	Exemples d'analyses psychométriques <ul style="list-style-type: none"> • Généralisabilité • Comparaison des résultats entre différents sous-groupes
Aspects externes	Liens entre le résultat de l'évaluation et les autres variables Validité convergente et divergente	Exemples d'analyses psychométriques <ul style="list-style-type: none"> • Corrélation des scores avec d'autres instruments de mesure • Comparaison des résultats avec d'autres sous-groupes • Régression multiple
Conséquences	Mesure des conséquences de l'évaluation Conséquences positives ou négatives, prévues ou imprévues	Analyses variées selon la conséquence mesurée (p. ex. taux de réussite, anxiété ou motivation)

Synthèse adaptée des travaux de AERA et al. 1999; Cook & Beckman, 2006; Downing, 2003; Messick, 1995; Streiner et al., 2015

2.3.3 Validation fondée sur l'accumulation des preuves

Dans sa conceptualisation de la validité, Kane (2006) propose une approche s'appuyant sur l'argument d'interprétation et l'argument de validité (*argument-based rationale*) (tableau 4) (Cook et al., 2015; Kane, 2013; Kane, 2006). Le premier argument, l'*argument d'interprétation*, vise à préciser comment les résultats de l'évaluation seront interprétés et quelle en sera l'utilisation (Kane, 2013; Kane, 2006; Schuwirth & van der Vleuten, 2012). L'argument d'interprétation comporte une chaîne de quatre inférences, lesquelles sont présentées dans les lignes qui suivent :

- 1) L'inférence de *l'observation au score* vise à traduire l'observation dans un score ou plusieurs scores. Différentes stratégies de validation qualitatives ou quantitatives sont utiles pour vérifier la qualité de cette inférence. Par exemple, la performance d'un apprenant qui complète un QCM sera traduite par une note. Le contenu du QCM peut être vérifié par un groupe d'experts ou encore par une analyse des items.
- 2) Dans l'inférence du *score observé au score universel*, la personne responsable de l'évaluation vérifie si les données recueillies précédemment reflètent l'ensemble des connaissances possibles dans le domaine évalué. Dans l'exemple du QCM explicité ci-dessus, la personne responsable de l'évaluation vérifie si les questions sélectionnées (p. ex. 50 questions) sont représentatives du domaine évalué (p. ex. la cardiologie). Des stratégies qualitatives et quantitatives peuvent être utilisées comme la triangulation des données ou encore la théorie de réponses aux items.
- 3) Par l'inférence du *score universel au niveau d'atteinte du domaine*, la personne responsable de l'évaluation utilise le score pour obtenir un portrait de la performance de l'apprenant dans le monde réel. Des stratégies de validation qualitatives (entrevues avec différentes parties prenantes) et quantitatives (p. ex. preuves de validité convergente ou divergente) sont possibles. Ainsi, dans l'exemple précédent, le niveau de connaissance que possède l'apprenant en cardiologie devrait être transférable dans le monde réel avec un patient (p. ex. effets secondaires de la médication, analyse d'un électrocardiogramme). Des entrevues avec différentes parties prenantes permettraient de vérifier si le QCM est pertinent pour préparer les apprenants à prendre soin de patients sur une unité de cardiologie.

- 4) L'inférence du *niveau d'atteinte du domaine aux décisions prises* se traduit par la prise de décision ou l'action posée à la suite de l'ensemble des inférences. Les conséquences découlant de ces décisions sont mesurées par des stratégies de validation qualitatives et quantitatives. Aussi, les personnes responsables de l'évaluation ont besoin de davantage d'informations sur la performance de l'apprenant avant de poser un jugement final (p. ex réussite ou échec). Des entrevues avec les apprenants pourraient renseigner sur les conséquences positives et négatives, prévues ou imprévues du QCM en cardiologie.

Le deuxième argument, l'*argument de validité*, a pour objectif d'examiner la qualité des preuves recueillies afin d'appuyer la décision découlant de l'interprétation des résultats (Cook et al., 2015; Kane, 2006). Dans la chaîne d'inférences, les preuves recueillies sont multiples. La qualité de chacune des preuves doit être examinée dans son ensemble (Pennaforde & Loye, 2017). L'importance d'une preuve peut varier selon le but et le contexte de l'évaluation. Par exemple, si l'évaluation est formative, l'inférence de généralisation (du score observé au score universel) peut être de moindre importance (Cook et al., 2015).

Tableau 4
Validation fondée sur l'accumulation des preuves

Argument	Description
Argument d'interprétation	Interprétation des résultats de l'évaluation Utilisation des résultats de l'évaluation
	Chaîne d'inférences
	Exemples de stratégies de validation
<i>De l'observation au score (score)</i> Traduction d'une observation par un ou plusieurs scores	Stratégies qualitatives Jugement d'experts Qualité des données narratives (p. ex. authenticité, richesse) Stratégies quantitatives Accord interjuges Indice de difficulté des items
<i>Du score observé au score universel (généralisation)</i> Représentation du domaine évalué	Stratégies qualitatives Jugement d'experts Triangulation des données Stratégies quantitatives Généralisabilité Théorie de réponses aux items
<i>Du score universel au niveau d'atteinte du domaine (extrapolation)</i> Reflet de la performance de l'apprenant dans le monde réel	Stratégies qualitatives Entrevues avec parties prenantes (p. ex. pertinence et authenticité de l'évaluation) Stratégies quantitatives Méthode multitraits-multiméthodes Validité convergente et divergente Analyse factorielle
<i>Du niveau d'atteinte du domaine aux décisions prises (décision)</i> Prise de décision ou action à la suite des inférences Réflexion sur la nécessité d'obtenir davantage d'informations sur la performance de l'apprenant	Stratégies qualitatives Entrevues avec parties prenantes (p. ex. mesure des conséquences positives/négatives) Stratégies quantitatives Mesures longitudinales des conséquences (p. ex. taux de réussite à l'examen de certification, taux de décrochage scolaire)
Argument de validité	Examine la cohérence de l'argument interprétatif et la crédibilité de ses conclusions

Synthèse adaptée des travaux de Cook et al., 2015; Kane, 2013; Kane, 2006; Schuwirth & van der Vleuten, 2012

2.3.4 Critiques des théories modernes

Cizek (2016) mentionne que le concept de validité doit évoluer et que nous devons aller au-delà des travaux de Messick et de Kane. Il rejette la vision proposée par Messick en précisant, entre autres, qu'il est difficile de réunir l'ensemble des preuves de validité identifiées. Cizek et al. (2016a; 2018) affirment que la justification de l'utilisation d'un test (évaluation) et les inférences d'une évaluation ne peuvent pas être combinés et nécessitent la collecte de preuves distinctes.

La théorie de la validité unifiée de Messick, (qui documente cinq ou six preuves de validité selon la version adoptée), reprise dans le document *Standards for Educational and Psychological Testing* (AERA et al., 1999; AERA et al., 2014), ne semble pas être fréquemment utilisée dans les recherches en pédagogie des sciences de la santé (Cook et al., 2013; Cook et al., 2014) ou dans le domaine de la mesure et évaluation (Cizek, Bowen, & Church, 2010). Par exemple, la preuve de conséquences est rarement documentée par les auteurs qui décrivent leur processus de validation (Cizek, Bowen, & Church, 2010; Cook & Lineberry, 2016; Cook et al., 2014; Labbé, Young, & Nguyen, 2018). Plusieurs hypothèses peuvent expliquer cette absence telles qu'une mauvaise compréhension de la preuve de conséquences (Cook & Lineberry, 2016) ou les efforts et le temps exigés pour mesurer cette preuve, souvent de manière longitudinale (Cizek et al., 2010). Les avis sont partagés quant à la preuve de conséquences. Pour certains auteurs (p. ex. Popham, 1997; ou Shepard, 1997), la preuve de conséquences n'est pas considérée comme une preuve de validité (Cizek, 2012; Cook & Lineberry, 2016; Downing, 2003; Hubley & Zumbo, 2011; Popham, 1997; Shepard, 1997; Sireci & Sukin, 2013). Popham (1997) et Shepard (1997) accordent de l'importance à la mesure des conséquences et précisent qu'elle doit être documentée par les personnes responsables de l'évaluation. Toutefois, cette preuve de validité n'indique pas si l'outil mesure bien ce qu'il devrait mesurer.

En bref, les processus de validation les plus fréquemment utilisés pour vérifier la qualité de l'évaluation des apprentissages sont : 1) la théorie unifiée de la validité de Messick (1995) et 2) la validation fondée sur l'accumulation des preuves de Kane (2006). Dans les écrits scientifiques répertoriés, des auteurs explorent la possibilité d'appliquer les théories

modernes de la validité pour mesurer la qualité de nouvelles pratiques évaluatives dans le contexte d'une APC (Cook & Beckman, 2006; Cook, Brydges, Ginsburg, & Hatala, 2015; Downing, 2003; Hodges, 2013; Lockyer et al., 2017; Schuwirth & van der Vleuten, 2006; Timmerman & Dijkstra, 2017; van der Vleuten et al., 2012). Toutefois, à notre connaissance, les chercheurs sont toujours à la recherche de solutions pour améliorer la qualité de l'interprétation des résultats dans le contexte de nouvelles pratiques évaluatives.

2.4 Écart entre les pratiques évaluatives et les processus de validation existants

Dans la première section de la recension des écrits, nous avons décrit les nouvelles pratiques évaluatives. Nous avons pisté le lecteur sur différentes problématiques retrouvées dans les écrits scientifiques. Dans la deuxième section, nous avons explicité les processus de validation mis de l'avant dans les théories modernes de la validité. Les nouvelles pratiques évaluatives semblent être élaborées davantage selon une perspective constructiviste. Or, les approches psychométriques existantes sont fondées sur une perspective postpositiviste. Cette divergence entre les deux pourraient justifier les insatisfactions exprimées par les chercheurs puisque les paradigmes représentent une manière différente de voir le monde. Nous souhaitons maintenant porter un regard sur l'écart entre les pratiques évaluatives et les processus de validation existants et décrire comment les chercheurs en pédagogie des sciences de la santé envisagent de les réconcilier.

Parmi les réflexions retrouvées dans les écrits, des auteurs ont exploré la possibilité d'utiliser les théories modernes de la validité pour mesurer la qualité des nouvelles pratiques évaluatives (Cook & Beckman, 2006; Cook, Brydges et al., 2015; Downing, 2003; Schuwirth & van der Vleuten, 2012). De manière plus spécifique, Downing (2003) ainsi que Cook & Beckman (2006) ont examiné comment utiliser la théorie de la validité unifiée pour vérifier la qualité de différentes méthodes d'évaluation utilisées dans les programmes de médecine.

Les processus de validation actuels (AERA et al., 2014) sont principalement élaborés pour mesurer la qualité d'une méthode d'évaluation prise individuellement et semblent difficilement transférables pour mesurer la qualité d'un ensemble de stratégies intégrées dans une évaluation programmatique (Schuwirth & van der Vleuten, 2006; Timmerman &

Dijkstra, 2017; van der Vleuten & Dannefer, 2012). Pour cette raison, des chercheurs ont examiné la possibilité d'utiliser la théorie de Kane dans le contexte de l'évaluation programmatique (Cook et al., 2015; Schuwirth & van der Vleuten, 2012). L'utilisation de critères de rigueur de recherche qualitative semble être une piste de solution pour appuyer les inférences provenant d'une évaluation narrative (Cook et al., 2016), d'une évaluation en milieu de travail (Govaerts & van der Vleuten, 2013) et de l'évaluation programmatique (Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2012).

La quantité de données recueillies dans un programme d'évaluation peut sembler énorme pour les personnes responsables de l'évaluation (Norman et al., 2014). L'interprétation de l'ensemble des données et l'ajout de données narratives provenant de l'observation d'évaluateurs nécessitent de revoir les processus de validation (Hodges, 2013; Lockyer et al., 2017). À cet effet, un échange regroupant les acteurs impliqués peut s'avérer nécessaire pour effectuer la synthèse des données recueillies et prendre une décision quant à la performance d'un apprenant (Lockyer et al., 2017).

Bien que les processus de validation soient continuellement révisés en fonction de l'évolution des pratiques évaluatives, bon nombre d'auteurs ont exprimé leur inconfort concernant l'utilisation des processus existants (Berendonk et al., 2013; Dijkstra et al., 2010; Gingerich, Kogan, et al., 2014; Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010; Hodges, 2013; Holmboe et al., 2017; Schuwirth & van der Vleuten, 2006; Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2012). Plus spécifiquement, les preuves de validation quantitatives (c.-à-d. analyses psychométriques) axées sur la qualité de la mesure ne semblent pas suffisantes et pourraient même engendrer des inférences inexactes (Govaerts & van der Vleuten, 2013). Schuwirth et van der Vleuten (2006) indiquent que :

Particularly in the light of current developments in assessment, we think that the strict psychometric model is no longer sustainable and we would like to make a plea for a major revision of the statistical concepts and approaches to assessment (Schuwirth & van der Vleuten, 2006, p. 296).

Conscientes de l'écart entre les nouvelles pratiques évaluatives et les processus de validation existants, les personnes engagées dans la recherche en pédagogie des sciences de la santé expriment le besoin d'explorer d'autres avenues (Hawkins et al., 2015; Norcini & McKinley,

2007; van der Vleuten et al., 2010). Par exemple, Govaerts et van der Vleuten mentionnent que :

We argue that meaningfulness and appropriateness of current validity evidence can be called into question and that we need alternative strategies to assessment and validity inquiry that build on current theories of learning and performance in complex and dynamic workplace settings (Govaerts & van der Vleuten, 2013, p. 1164).

Par ailleurs, dans une analyse de discours, St-Onge et al. (2017) ont identifié trois différentes conceptualisations de la validité, à savoir : 1) la validité comme une caractéristique d'un test; 2) la validité comme une accumulation de preuves et 3) la validité en tant qu'impératif social. Cette dernière conceptualisation (la validité en tant qu'impératif social) rassemble les insatisfactions vécues face aux processus de validation existants dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé. Les écrits scientifiques répertoriés par St-Onge et al. (2017) ne décrivent pas de manière détaillée ce qu'est la validité en tant qu'impératif social ni comment la mettre en application dans la pratique. La conceptualisation émergente de la validité en tant qu'impératif social se caractérise par le recours à des preuves de validation qualitatives pour compléter les preuves de validation quantitatives, à la vision holistique et englobante de la validité et à la considération des conséquences de l'évaluation sur l'apprenant et sur la société (St-Onge et al., 2017). La validité en tant qu'impératif social pourrait permettre une validation plus appropriée des pratiques évaluatives actuellement privilégiées dans les programmes professionnalisants en sciences de la santé.

En résumé, la vérification de la qualité de l'évaluation des apprentissages représente une préoccupation importante pour les personnes qui en sont responsables. Les questionnements et les insatisfactions recueillis dans la littérature proviennent principalement d'articles scientifiques (documents de références, articles théoriques, éditoriaux, commentaires). Les insatisfactions exprimées dans les écrits scientifiques concernant les limites des processus de validation existants sont à l'origine d'une conceptualisation émergente de la validité : la validité en tant qu'impératif social (St-Onge et al., 2017). La validité en tant qu'impératif social se caractérise par une vision plus holistique de la validité ainsi qu'une considération marquée pour les conséquences de l'évaluation sur l'apprenant et sur la société.

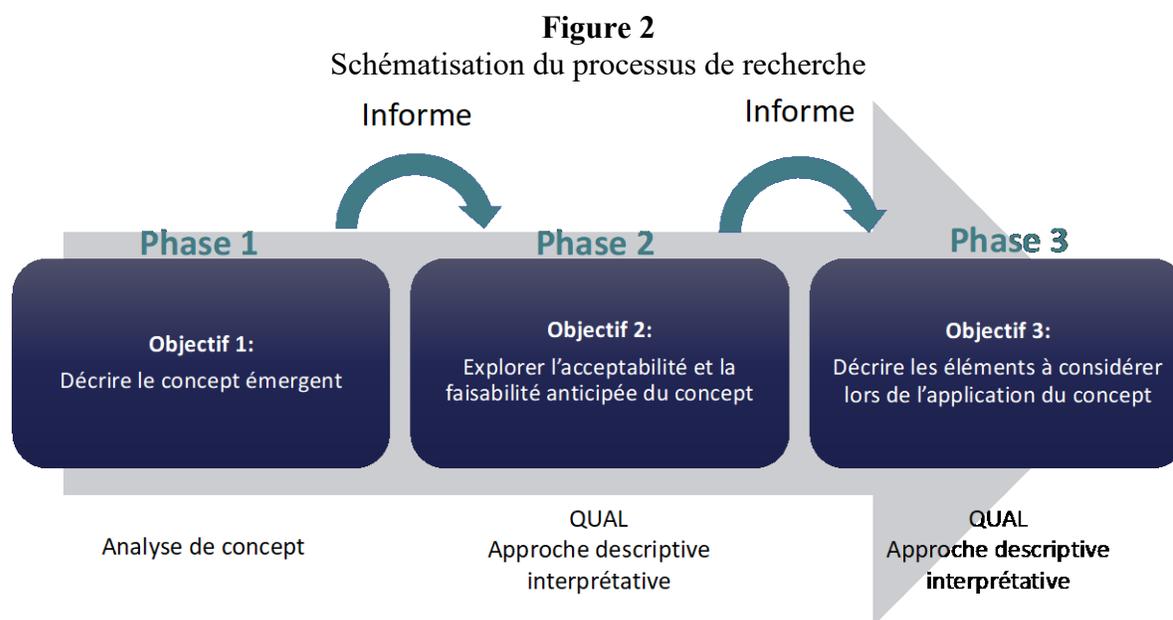
3. MÉTHODE

Dans le chapitre 3, nous présenterons une schématisation de l'ensemble du processus de recherche et une brève description de la méthode spécifique pour chacune des trois phases de l'étude. Dans les chapitres 4, 5 et 6, nous décrirons la méthode détaillée pour chacune des phases à l'intérieur de trois articles.

3.1 Devis

Nous avons utilisé une recherche exploratoire en trois phases pour mieux comprendre le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé. La figure 2 illustre visuellement les méthodes utilisées pour chacune des phases afin de bien situer le lecteur. Dans la phase 1, une analyse de concept visait à décrire le concept émergent de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages (objectif 1). Dans la phase 2, un devis descriptif interprétatif a permis d'explorer l'acceptabilité et la faisabilité anticipée du concept de validité en tant qu'impératif social (objectif 2). Des entrevues individuelles et des groupes de discussion focalisée ont été menés auprès des différentes parties prenantes. Dans la phase 3, un devis descriptif interprétatif a été utile pour décrire les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social (objectif 3).

La recherche en trois phases se décline de façon séquentielle, chacune des phases informant la suivante. Cet enchaînement nous a permis de voir comment la validité en tant qu'impératif social est perçue selon différentes perspectives, c'est-à-dire selon les auteurs d'écrits scientifiques et selon différentes parties prenantes. La combinaison de plus d'une méthode de recherche a permis une exploration en profondeur du sujet.



3.2 Cadre paradigmatique

Actuellement, dans les écrits scientifiques, la validité de l'évaluation des apprentissages en tant qu'impératif social a un niveau d'abstraction très élevé, ce qui nécessite une approche de recherche inductive et exploratoire, comme le suggèrent LoBiondo-Wood, Haber, Cameron, & Singh (2018) et Varpio, Paradis, Uijtdehaage, & Young (sous presse). Selon des auteurs, il est possible de s'appuyer sur un cadre méthodologique pour guider l'ensemble de la recherche (LoBiondo-Wood et al., 2018) ou encore sur des notions philosophiques (Fortin & Gagnon, 2016). Le caractère émergent du concept à l'étude et la nature exploratoire de la démarche justifient la décision de fonder la recherche sur un cadre paradigmatique. Ainsi, dans la présente étude, toutes les décisions sont influencées par le paradigme constructiviste (LoBiondo-Wood et al., 2018; Mertens, 2019), c'est pourquoi nous avons choisi de parler d'un cadre paradigmatique.

Sur le plan ontologique, nous avons considéré différentes perceptions du concept de la validité en tant qu'impératif social, tant dans les écrits consultés lors de l'analyse de concept que lors de la consultation des participants à la recherche. Sur le plan épistémologique, lors des entrevues de la phase 2 de la recherche (section 3.4), une attention a été portée par l'animatrice afin de créer un climat de confiance. Des questions ouvertes ont permis de tenir compte de la subjectivité des participants. Sur le plan méthodologique, le paradigme

constructiviste a guidé nos décisions (Mertens, 2019), notamment en ce qui concerne le choix du devis ainsi que les décisions relatives à collecte et à l'analyse des données qualitatives et ce, pour toutes les phases de la recherche.

Sur le plan axiologique, la recherche est influencée par les valeurs des membres de l'équipe de recherche (LoBiondo-Wood et al., 2018; Mertens, 2019; Woo, 2019). En cohérence avec le paradigme constructiviste, Thorne (2016) suggère aux chercheurs de réfléchir à l'influence possible de leur discipline et de leur expérience sur l'analyse et l'interprétation des résultats avant de débiter la recherche. Notre équipe est composée de quatre personnes concernées par la qualité de l'évaluation des apprentissages. L'étudiante possède une formation initiale en sciences infirmières et enseigne au baccalauréat depuis plusieurs années. Elle élabore et administre ses propres méthodes d'évaluation des apprentissages. Les trois directrices sont professeures et chercheuses dans le domaine de la recherche qualitative et en pédagogie des sciences de la santé. Afin d'assurer la distanciation nécessaire à la prise en compte de la réalité des personnes participant à la recherche, nous avons discuté de notre expérience et de notre vision de la validité et de l'évaluation des apprentissages avant d'amorcer la collecte de données et, de façon régulière, au cours de l'analyse des données.

Bien que les théories modernes de la validité (Messick, 1995 et Kane, 2006) et les concepts d'acceptabilité et de faisabilité (Sidani et Braden, 2011) soient utilisés dans la présente recherche, ils ne peuvent être considérés comme cadre théorique ou cadre conceptuel. Les théories de Messick (1995) et de Kane (2006) explicitées dans le chapitre 2 servent de point de départ pour comprendre les processus de validation existants. Toutefois, les théories modernes de la validité ne peuvent servir d'ancrage pour l'ensemble de la recherche puisque ces théories se distinguent du concept de validité en tant qu'impératif social. Les concepts d'acceptabilité et de faisabilité définis par Sidani et Braden (2011) ont orienté les questions utilisées lors des entrevues et ont guidé l'analyse des données (thèmes prédéterminés) (section 5.2.1).

3.3 Phase 1 : Description du concept émergent

Une analyse de concept selon la méthode développée par Rodgers (1989) a permis de décrire le concept émergent de la validité de l'évaluation des apprentissages en tant qu'impératif

social, soit les antécédents, les attributs et les conséquents de ce concept (objectif 1). Dans cette section, nous décrirons brièvement la méthode utilisée lors de la phase 1 de la recherche doctorale. La méthode est détaillée dans un premier article disponible au chapitre 4.

Au même titre que l'ensemble des méthodes de recherche habituellement reconnues, l'analyse de concept contribue au développement des connaissances scientifiques par une démarche rigoureuse (Baldwin, 2008; McEwen & Wills, 2019). L'analyse de concept va au-delà de la synthèse d'écrits scientifiques, elle exige une analyse critique de documents et ses résultats permettent de produire de nouvelles connaissances (Morse, 2000). Dans une analyse de concept, la stratégie d'analyse des données narratives des documents s'apparente à celle utilisée en recherche qualitative.

L'analyse de concept s'avère une méthode pertinente pour identifier, clarifier et raffiner un concept inexploré (Hodges, Kuper, & Reeves, 2008; Naumanen-Tuomela, 2001; Rodgers, 1989; Rodgers & Knafl, 2000; Tofthagen & Fagerstrøm, 2010; Wills & McEwen, 2011). Cette première phase représente une étape cruciale dans le processus du développement des connaissances (Rodgers & Knafl, 2000) au sujet du concept de la validité en tant qu'impératif social et pour la poursuite des phases subséquentes de la présente recherche. Dans une perspective constructiviste, nous avons exploré les perceptions de différents auteurs afin de mieux comprendre le concept émergent.

Il existe plusieurs méthodes d'analyse de concept, notamment celles développées par Walker et Avant (2019) et Wilson (1963). L'analyse de concept est une méthode reconnue, rigoureuse et structurée (Hupcey, Morse, Lenz, & Tazón, 1996; Rodgers & Knafl, 2000). Nous avons privilégié la méthode de Rodgers (1989) qui se démarque par l'importance accordée au contexte ainsi qu'au caractère évolutif du concept (Hupcey et al., 1996; Wills & McEwen, 2011). En raison de l'émergence du concept de la validité en tant qu'impératif social dans la littérature, il nous semblait important de considérer son évolution future.

L'analyse de concept selon Rodgers comprend six étapes (tableau 5). La description détaillée des quatre premières étapes réalisées se retrouve dans l'article 1, présenté au chapitre 4. Nous

avons considéré qu'il était trop tôt pour effectuer les étapes 5 et 6, soit d'identifier un exemple, des implications et des hypothèses, en raison de la nouveauté du concept et de l'application du concept encore inexplorée à ce jour.

Tableau 5
Analyse de concept selon Rodgers

Étapes	Activités selon Rodgers
1	Identifier le concept d'intérêt
2	Identifier et sélectionner les paramètres de l'échantillon (documents)
3	Collecter les données
4	Analyser les données
5	Identifier un exemple
6	Identifier les implications et les hypothèses

Rodgers (1989) et Rodgers & Knafl (2000)

3.4 Phase 2 : Acceptabilité et faisabilité anticipée du concept

Après avoir décrit la validité en tant qu'impératif social par une analyse de concept, nous avons exploré l'acceptabilité et la faisabilité anticipée du concept (objectif 2) en utilisant une approche descriptive interprétative (Thorne, 2016). Dans cette section, nous présentons un aperçu de la méthode de la phase 2. Au chapitre 5, nous explicitons davantage la méthode à l'intérieur d'un deuxième article et de compléments méthodologiques.

L'approche descriptive interprétative a été développée pour répondre aux besoins des disciplines appliquées (p. ex. sciences de la santé ou en sciences de l'éducation). Cette approche qualitative propose une démarche rigoureuse visant à comprendre et à expliquer le point de vue des participants ainsi que leur interprétation du phénomène étudié (Thorne, 2016). Elle facilite la production de connaissances significatives pouvant être interprétées et utilisées dans la pratique (Hunt, 2009; Thorne, 2016; Thorne, Reimer Kirkham & O'Flynn-Magee, 2004). Dans la phase 2 de la recherche, nous souhaitons explorer l'acceptabilité d'un concept émergent auprès des acteurs concernés et connaître leur point de vue au sujet des

barrières et des éléments facilitant son application dans la pratique des enseignants en sciences de la santé.

En cohérence avec l'approche descriptive interprétative de Thorne (2016), nous avons mené des entrevues individuelles et des entrevues de groupes de discussion focalisée auprès de deux profils de participants distincts, soit 1) le personnel enseignant et les membres de comités d'évaluation des apprentissages et 2) les personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada. Nous avons utilisé principalement une stratégie d'échantillonnage à choix raisonné pour recruter des participants jouant un rôle différent dans le contexte de l'évaluation des apprentissages. Le personnel enseignant et les membres de comités d'évaluation des apprentissages développent, administrent et monitorent les méthodes d'évaluation des apprentissages. Les personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada contribuent à l'avancement des connaissances dans le domaine de la validité ou de l'évaluation des apprentissages. Bien que le nombre exact de participants soit difficile à déterminer à l'avance (Malterud, Siersma, & Guassora, 2016), nous avons initialement planifié rencontrer entre 25 et 35 personnes présentant des caractéristiques diversifiées, afin d'optimiser la richesse des données. Ce nombre a été déterminé afin 1) de former des groupes de discussion focalisée composés de cinq à huit membres du personnel enseignant et de comités d'évaluation dans chacune des quatre université ciblées (Université de Montréal, Université Laval, Université de Sherbrooke ou Université McGill) et 2) d'interviewer cinq à huit personnes engagées dans la recherche au Canada. Cette représentativité des profils de participants, des milieux d'enseignement et des disciplines avait pour but d'explorer plusieurs points de vue sur le phénomène d'intérêt.

Les groupes de discussion focalisée représentaient notre premier choix pour rencontrer le personnel enseignant et les membres de comités d'évaluation des apprentissages. Nous souhaitons regrouper les personnes concernées par la qualité de l'évaluation des apprentissages autour d'une discussion. Nous avons formé des groupes homogènes (Morgan, 2019), c'est-à-dire en effectuant les rencontres auprès des utilisateurs œuvrant dans la même université. Les discussions ont été enrichies par l'interaction entre les participants (Fortin & Gagnon, 2016; Krueger & Casey, 2015; Morgan, 2019).

Pour collecter les données auprès des personnes engagées dans la recherche au Canada, nous avons opté pour des entrevues téléphoniques individuelles vu la dispersion de ces personnes

dans différentes universités et provinces canadiennes. L'entrevue individuelle semi-dirigée permet d'explorer en profondeur le sujet d'étude auprès de la personne interviewée (Brinkman & Kvale, 2015; Denzin & Lincoln, 2018; Fortin & Gagnon, 2016; Thorne, 2016). Le tableau 6 résume les décisions méthodologiques de la phase 2 pour chacun des profils de participants.

Tableau 6
Résumé des décisions méthodologiques de la phase 2

Profil de participant	Personnel enseignant et membres de comités d'évaluation des apprentissages	Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada
Devis	Approche descriptive interprétative	Approche descriptive interprétative
Critères d'admissibilité	Critères d'inclusion <ul style="list-style-type: none"> • Professeurs, chargés de cours ou membres de comités d'évaluation depuis plus d'un an • Implication dans les programmes de sciences infirmières, médecine, physiothérapie et ergothérapie • Employés de l'Université McGill, Sherbrooke, Montréal ou Laval • Contribution à la conception, à la validation ou au monitoring de méthodes d'évaluation des apprentissages Critères d'exclusion <ul style="list-style-type: none"> • Les personnes qui n'enseignent que quelques heures dans un cours, les tuteurs, les moniteurs ou les superviseurs de stage 	Critères d'inclusion <ul style="list-style-type: none"> • Professeurs, chercheurs, décideurs • Contributions significatives dans le domaine de la validité ou de l'évaluation des apprentissages en pédagogie des sciences de la santé au Canada Critères d'exclusion <ul style="list-style-type: none"> • Aucun
Échantillon	Par choix raisonné	Par choix raisonné Boule de neige
Recrutement	Par courriel via les directeurs de programmes	Par courriel
Durée de la collecte des données	Octobre à novembre 2016	Janvier à juillet 2017
Méthode de collectes des données	Groupes de discussion focalisée et entrevues individuelles en présentiel menés par l'étudiante Entrevues enregistrées d'une durée de 60 à 90 minutes	Entrevues individuelles par téléphone menées par deux professionnelles de recherche Entrevues enregistrées d'une durée de 60 à 90 minutes
Outils de collectes de données	Questionnaire sociodémographique Guide d'entrevue semi-structuré	Questionnaire sociodémographique Guide d'entrevue semi-structuré

3.4.1 *Analyse des données*

Nous avons utilisé des statistiques descriptives (nombre, proportion, moyenne et écart type, selon le cas) pour analyser les données quantitatives provenant des questionnaires sociodémographiques.

La méthode utilisée pour analyser les données qualitatives des deux profils de participants est similaire. Ainsi, le contenu de l'enregistrement des entrevues (de groupe et individuelles auprès des deux profils de participants) a été transcrit afin de faciliter l'analyse qualitative de l'ensemble des données. Le contenu des notes prises par l'observatrice lors des groupes de discussion focalisée a été inséré à la transcription lorsque nécessaire. Nous nous sommes assurés de la qualité de la transcription des données par l'écoute de la bande audio. L'étudiante et une directrice ont procédé à l'analyse qualitative de l'ensemble des données et, par la suite, tous les membres de l'équipe ont discuté de l'interprétation des données jusqu'à l'atteinte d'un consensus. L'analyse des données a été guidée par la méthode de Miles, Huberman et Saldana (2014) qui consiste en trois activités d'analyse simultanée réalisées de manière itérative et concomitante, à savoir : 1) la condensation des données, 2) la présentation des données sous forme de tableaux ou schémas et 3) la vérification des conclusions. Le choix de la méthode d'analyse des données qualitatives était cohérent avec l'approche descriptive interprétative (Thorne, 2016), notamment par la prise en compte du point de vue des participants à propos du concept de validité en tant qu'impératif social. Miles, Huberman et Saldana (2014) recommandent de mener la collecte des données et l'analyse des données de manière simultanée. Cette stratégie permet d'améliorer la collecte des données futures en ajustant, par exemple, le guide d'entrevue. Dans la présente recherche, il n'a pas été possible de mener la collecte des données et l'analyse des données de manière concomitante pour un même profil de participants. Par exemple, pour le personnel enseignant et les membres de comités d'évaluation des apprentissages, les entrevues de groupes et individuelles se sont déroulées la même journée en raison de la disponibilité des participants ou de la distance entre les différentes universités. Pour atteindre un niveau satisfaisant de saturation des données et ajuster le guide d'entrevue, l'étudiante et une directrice ont écouté les enregistrements et rédigé un résumé d'une page après chaque entrevue. Elles ont ensuite

discuté du contenu avant la prochaine entrevue dans l'optique de raffiner le déroulement des entrevues subséquentes.

La première activité d'analyse (condensation des données) consistait à analyser les verbatims des entrevues. La grille d'analyse initiale comportait des codes prédéterminés fondés sur les définitions de l'acceptabilité et de la faisabilité proposées par Sidani et Braden (2011). L'étudiante et une directrice ont lu les transcriptions des entrevues et ont procédé à l'analyse des données de façon inductive par l'ajout de codes émergents. Les codes prédéterminés et des exemples de codes émergents figurent au tableau 7. Cette activité d'analyse a été effectuée à l'aide du logiciel Dedoose (2018).

Tableau 7
Codes prédéterminés et exemples de codes émergents

Codes prédéterminés Inspirés de Sidani et Braden, 2011	Exemples de codes émergents
Acceptabilité <ul style="list-style-type: none"> • Pertinence du concept de validité en tant qu'impératif social • Commodité • Efficacité • Risques associés • Adhésion à cette vision de la validité 	<ul style="list-style-type: none"> • Compréhension de l'attribut ou du concept (p. ex. signification du mot société, description d'un attribut difficile à comprendre pour le participant) • Conscientisation à l'importance de la validité (p. ex. observations des participants concernant les conséquences pour l'apprenant ou la société)
Faisabilité <ul style="list-style-type: none"> • Disponibilité et qualité des personnes évaluatrices • Formation des personnes évaluatrices • Ressources nécessaires • Contexte • Fidélité • Portée 	<ul style="list-style-type: none"> • Difficultés envisagées en lien avec l'utilisation du concept (p. ex. pression des organismes accréditeurs, barrières administratives de l'établissement d'enseignement) • Éléments facilitant l'utilisation du concept (p. ex. soutien facultaire, création d'un comité)

Simultanément, la deuxième activité d'analyse (présentation des données sous forme de tableaux ou schémas) a favorisé l'interprétation des données (Gallagher, 2014; Miles et al., 2014). L'utilisation de matrices conceptuelles (Miles et al., 2014) nous a permis d'organiser

et de visualiser les données recueillies pendant les entrevues de la phase 2. Nous avons commencé l'analyse en explorant les réponses des participants quant à leur vision de l'acceptabilité et de la faisabilité anticipée pour chacun des quatre attributs du concept de la validité en tant qu'impératif social. À la suite de cette première analyse, nous avons produit quatre tableaux, un pour chacun des attributs. Dans le tableau 8, nous illustrons un extrait d'une matrice conceptuelle élaborée pour l'attribut 1. Les réponses des participants ont été classées sous trois différentes catégories, soit 1) perception positive (c.-à-d. l'attribut est acceptable ou applicable); 2) perception positive avec réserves (c.-à-d. l'attribut est acceptable ou applicable, mais des réserves sont émises) et 3) perception négative (c.-à-d. l'attribut n'est pas acceptable ni applicable).

Tableau 8
Extrait d'une matrice conceptuelle portant sur l'attribut 1 (phase 2)

Attribut 1			
Preuves de validité jugées crédibles par la société			
	Perception positive	Perception positive avec réserves	Perception négative
Acceptabilité	La validité doit être accessible aux différentes parties prenantes	Acceptabilité selon le contexte Questionnement sur la nécessité de fournir des preuves à la société Confusion dans l'utilisation du terme « société »	Difficile de décrire les attentes de la société à propos des preuves de validité nécessaires
Faisabilité	Les préoccupations actuelles des instances professionnelles peuvent faciliter l'application de l'attribut		Défis potentiels de l'application de l'attribut

Dans un cinquième tableau, nous avons analysé les données concernant la perception des participants sur le concept de validité en tant qu'impératif social dans son ensemble. Des cinq tableaux résultant de l'analyse, nous avons examiné les divergences (résultats uniques à un

tableau), les similitudes (chevauchement entre les tableaux) et les liens (entre les différents tableaux) afin de tirer des conclusions.

La troisième activité d'analyse (vérification des conclusions) (Gallagher, 2014; Miles et al., 2014) a été effectuée principalement à l'aide de discussions et de réflexions régulières au sein de l'équipe. L'étudiante a analysé toutes les entrevues et une directrice les a coanalysées. Nous avons rédigé des mémos dans le logiciel Dedoose au sujet des questions méthodologiques et analytiques. Ces questions (p. ex. décision de regrouper les catégories similaires ou non) étaient discutées au sein de l'équipe. En cas de divergences d'opinions, nous avons réexaminé les données brutes et avons discuté jusqu'à l'obtention d'un consensus. Cette réflexion, tout au long du processus d'analyse, a permis d'améliorer l'interprétation des résultats. Nous avons vérifié la justesse de notre interprétation auprès des personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada. Nous avons transmis le résumé de notre interprétation aux participants, ce qui leur a permis de se prononcer par courriel sur l'exactitude de leurs propos exprimés lors des entrevues que nous avons réalisées. Des modifications ont été apportées au besoin. Cette étape nous semblait essentielle puisque les entrevues ont été menées par une tierce personne (professionnelle de recherche), bénéficiant de la supervision de l'étudiante-chercheuse. Nous n'avons pas effectué de validation auprès du personnel enseignant et des membres de comités d'évaluation des apprentissages. La doctorante a interrogé ce profil de participants en présentiel, permettant de valider la compréhension des propos pendant les entrevues. Par exemple, les personnes ont été invitées à clarifier ce qu'elles aimaient ou ce qu'elles n'aimaient pas à propos d'un attribut.

3.4.2 Considérations éthiques

Les considérations éthiques concernent la phase 2 de la recherche. Nous avons obtenu l'approbation du Comité d'Éthique et de la Recherche en Éducation et en Sciences Sociales (CÉR-ESS) de l'Université de Sherbrooke le 22 juin 2016 (annexe 1). Nous avons également obtenu l'approbation éthique des autres universités où se déroulait la recherche (annexes 2, 3 et 4). Les participants ont reçu les informations concernant la recherche avant leur participation. Au besoin, nous avons répondu aux questions des personnes potentiellement

intéressées. La participation à l'étude était volontaire et les participants ont pu se retirer de l'étude en tout temps. Avant l'entrevue, le formulaire de consentement a été rempli sur papier (personnel enseignant et membres de comités d'évaluation des apprentissages) ou en ligne (personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada).

Pour les personnes participant aux entrevues de groupe, nous avons sollicité leur engagement moral à préserver la confidentialité du contenu des échanges. L'étudiante a assuré la confidentialité des données recueillies et anonymisé les transcriptions. Le matériel audionumérique et électronique est protégé par un mot de passe et conservé dans une base de données sur un disque dur de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke (SMUS). Les questionnaires sociodémographiques ainsi que les transcriptions recueillies en format papier sont conservés dans un classeur fermé à clé dans le bureau de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la SMUS.

3.5 Phase 3 : Éléments à considérer lors de la mise en œuvre du concept de validité en tant qu'impératif social

Au cours de la phase 3, nous avons décrit les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social (objectif 3). Pour répondre à cet objectif, nous avons intégré les résultats de la phase 1 (analyse de concept) et de la phase 2 (entrevues). Dans la prochaine section, nous décrirons brièvement comment nous avons procédé. Au chapitre 6, un article et des données complémentaires servent à expliciter la méthode utilisée.

3.5.1 Intégration des données

Dans la phase 3, l'objectif de l'intégration des données consistait à porter un regard au-delà des résultats des premières activités d'analyse afin d'obtenir une meilleure compréhension du sujet de recherche (Leavy, 2017). Plus spécifiquement, à partir d'une vision globale de tous les résultats réunis (phase 1 et phase 2), nous souhaitons dégager les éléments à considérer pour l'application de la validité en tant qu'impératif social. L'approche descriptive interprétative de Thorne (2016) et la méthode d'analyse des données qualitatives de Miles et

al. (2014) ont guidé l'intégration des données dans l'optique de répondre au troisième objectif de la recherche.

La première activité d'analyse (condensation des données) consistait à analyser les résultats provenant de l'analyse de concept (phase 1) et des entrevues (phase 2). Les résultats ont été rassemblés dans un seul document Excel sans égard à leur provenance. Au besoin, nous avons également ajouté de nouvelles données provenant de la littérature scientifique en éducation. Ces écrits ont été sélectionnés par boule de neige, selon les besoins d'approfondissement de certains sujets. Nous avons procédé de façon inductive, sans codes prédéterminés, avec l'objectif de rendre applicables les attributs du concept de validité en tant qu'impératif social. L'étudiante et une directrice ont procédé à l'analyse qualitative des données regroupées dans une perspective d'intégration des données. Il a été nécessaire de retourner fréquemment aux données brutes afin de confirmer notre compréhension tel que suggéré par Thorne (2016).

Dans le cadre de la deuxième activité d'analyse (présentation des données sous forme de tableaux ou schémas), l'utilisation de matrices conceptuelles (Miles et al., 2014) a facilité l'organisation des données. Nous avons utilisé des tableaux pour intégrer les différents éléments à considérer, ce qui nous a permis de déplacer, ordonner, organiser, replacer, déconstruire et reconstruire les matrices. Selon Thorne (2016), la transformation des données nécessite de réfléchir aux différentes relations entre les données et à la question de recherche. Nous nous sommes questionnés sur ce qui était visible (p. ex. similarités et divergences entre les attributs), mais également sur ce qui était invisible (p. ex. absence de détails sur certaines preuves de validité). Le tableau 9 illustre, par un exemple, l'intégration des données. Les différentes sources de données sont répertoriées dans la troisième colonne. Il s'agit des auteurs identifiés lors de la phase 1 (analyse de concept) ainsi que les verbatim obtenus lors de la phase 2 (entrevues individuelles et groupes de discussion focalisée). L'analyse des données provenant des phases 1 et 2 a généré des codes émergents détaillé dans la deuxième colonne. Ces codes émergents ont été reliés aux attributs du concept de validité en tant qu'impératif social dans la première colonne.

Tableau 9
Extrait d'une matrice conceptuelle (phase 3)

Attribut	Codes émergents	Sources des données
Validation intégrée au processus d'évaluation et à l'interprétation des résultats obtenus	<i>Validity build-in</i> (a priori)	<p><i>Écrits scientifiques :</i> (Baartman et al., 2013; Berendonk et al., 2013; Cook, Brydges, Ginsburg, & Hatala, 2015; Crossley, 2013; Dijkstra et al., 2010; Eva et al., 2015; Kasanda et al., 2013; Schuwirth & van der Vleuten, 2019; Shumway & Harden, 2003; Timmerman & Dijkstra, 2017; van der Vleuten et al., 2012; van der Vleuten et al., 2005; van der Vleuten et al., 2010)</p> <p><i>Entrevues :</i> N/A</p>
Décisions découlant des inférences		<p><i>Écrits scientifiques :</i> (Cook & Lineberry, 2016; Darling-Hammond & Snyder, 2000; DeLuca, 2011; Dijkstra et al., 2010; Hays, 2008; Kasanda, Mitonga, Vei, & Zimba, 2013; Kealey, 2010; Messick, 1995; Schuwirth & van der Vleuten, 2012; Schuwirth & van der Vleuten, 2011; van der Vleuten & Schuwirth, 2005; van der Vleuten et al., 2010)</p> <p><i>Entrevues :</i> Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada</p>
Importance des conséquences découlant de l'évaluation des apprentissages		<p><i>Écrits scientifiques :</i> (Cook et al., 2015; Cook & Lineberry, 2016; Dijkstra et al., 2010; Schuwirth & van der Vleuten, 2012; Schuwirth & van der Vleuten, 2011)</p> <p><i>Entrevues :</i> Personnel enseignant et membres de comités d'évaluation des apprentissages Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada</p>
Anticipation des conséquences		<p><i>Écrits scientifiques :</i> (Berendonk et al., 2013; Cook, et al., 2015; Cook & Lineberry, 2016; Dijkstra et al., 2010; Schuwirth & van der Vleuten, 2012)</p> <p><i>Entrevues :</i> Personnel enseignant et membres de comités d'évaluation des apprentissages</p>

La troisième activité d'analyse (vérification des conclusions) a nécessité des discussions régulières avec l'ensemble de l'équipe dans l'optique d'obtenir un consensus et de confirmer les conclusions. Ces rencontres ont permis de réajuster l'interprétation (transformation) des données et, ainsi, de répondre à l'objectif fixé.

La démarche utilisée pour intégrer les données est cohérente avec l'approche descriptive interprétative. Thorne (2016) accorde de l'importance à l'analyse avancée des données qu'elle appelle « transformation des données ». Selon cette approche, les chercheurs approfondissent les premières analyses et réfléchissent aux différentes relations entre les données et la question de recherche.

4. MÉTHODE, RÉSULTATS ET DISCUSSION DE LA PHASE 1 - ANALYSE DE CONCEPT

Dans le chapitre 4, nous présenterons la méthode, les résultats et la discussion d'une analyse de concept sous la forme d'un premier article. Ce volet correspond au premier objectif du projet doctoral qui vise à décrire le concept émergent de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages.

Article 1

**Validity as a social imperative for assessment
in health professions education: a concept analysis****Auteurs de l'article**

Marceau, Mélanie; Gallagher, Frances; Young, Meredith; St-Onge, Christina

Statut de l'article

Publié.

L'autorisation de reproduction à l'éditeur de la revue a été demandée.

Marceau, M., Gallagher, F., Young, M., & St-Onge, C. (2018). Validity as a social imperative for assessment in health professions education: a concept analysis. *Medical Education*, 52(6), 641–653. <https://doi.org/10.1111/medu.13574>

Avant-propos de l'article

Nous présentons dans cet article la méthode, les résultats et la discussion de l'objectif 1 de l'étude doctorale qui vise à décrire le concept émergent de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages.

Contribution de l'étudiante et des coauteurs

L'étudiante a réalisé l'analyse de concept en collaboration avec les coauteurs. Elle a élaboré les critères d'inclusion et d'exclusion de manière itérative et effectué la collecte des données dans les banques de données. Accompagnée par l'ensemble de l'équipe, l'étudiante a sélectionné les documents pertinents et a effectué l'analyse dans le logiciel NVivo. Les données qualitatives ont été coanalysées par l'étudiante et ses directrices. L'étudiante a rédigé le premier jet du manuscrit et a procédé au raffinement du contenu, qui a été bonifié par les réflexions et commentaires de l'ensemble des auteurs. Les directrices de recherche ont supervisé l'ensemble des travaux en plus de contribuer à la rédaction de l'article.

Title:

Validity as a social imperative for assessment in health professions education: a concept analysis

Authors

Mélanie Marceau, MSc, PhD candidate, Frances Gallagher, PhD, Meredith Young, PhD, and Christina St-Onge, PhD.

M. Marceau is a PhD candidate, Department of Nursing, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada.

F. Gallagher is Professor, Department of Nursing, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada.

M. Young is Associate professor, Department of Medicine and Institute for Health Sciences Education, Faculty of Medicine, McGill University, Montreal, Québec, Canada.

C. St-Onge is Professor, Department of Medicine, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada. She holds the *Paul Grand'Maison de la Société des médecins* Research Chair in Medical Education.

Résumé

Contexte:

L'évaluation des apprentissages peut entraîner de lourdes conséquences pour les futurs professionnels de la santé et pour la société. Par conséquent, il s'avère essentiel de vérifier la qualité de l'évaluation. Les théories modernes de la validité semblent inappropriées pour vérifier la qualité des pratiques évaluatives complexes, telles que l'évaluation authentique et l'évaluation programmatique. Ici, nous explorons le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé. Le concept de validité en tant qu'impératif social pourrait servir de cadre lors de l'examen de la qualité des pratiques évaluatives.

Méthode:

Nous avons effectué une analyse de concept à l'aide de la méthode évolutive de Rodgers afin de décrire le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé. Soutenues par une bibliothécaire académique, nous avons effectué une démarche de recherche documentaire dans plusieurs banques de données pour obtenir des documents publiés entre 1995 et 2016. Sur un total de 321 résumés, nous avons identifié 67 articles répondant à nos critères d'inclusion. L'étudiante et une directrice ont analysé les documents à l'aide d'une approche d'analyse des données qualitatives. Tous les membres de l'équipe ont discuté de l'interprétation des données jusqu'à l'atteinte d'un consensus.

Résultats:

Les caractéristiques du concept sont: (i) le recours à des preuves jugées crédibles aux yeux de la société pour documenter la qualité de l'évaluation des apprentissages; (ii) la validation intégrée au processus d'évaluation et à l'interprétation du score; (iii) l'utilisation de preuves de validité appuyant l'interprétation de la combinaison des résultats de l'évaluation, et (iv) la démonstration d'une utilisation justifiée de preuves mixtes (quantitatives et qualitatives) pour documenter la qualité de l'ensemble des stratégies de l'évaluation des apprentissages.

Conclusion:

Le concept émergent de la validité en tant qu'impératif social met en évidence certains éléments présents dans les théories modernes de la validité, tandis que certaines caractéristiques semblent uniques au contexte de la pédagogie des sciences de la santé et vont au-delà des processus de validation existants. L'étude reflète l'importance de considérer les préoccupations de la société lors de l'évaluation des apprentissages et tout au long du processus de validation. Le concept de validité en tant qu'impératif social pourrait constituer une perspective prometteuse pour examiner la qualité des pratiques évaluatives complexes.

Mots clés:

Validité, évaluation, pédagogie des sciences de la santé, analyse de concept

Abstract**Context:**

Assessment can have far-reaching consequences for future health care professionals and for society. Thus, it is essential to establish the quality of assessment. Few modern approaches to validity are well situated to ensure the quality of complex assessment approaches, such as authentic and programmatic assessments. Here, we explore and delineate the concept of validity as a social imperative in the context of assessment in health professions education (HPE) as a potential framework for examining the quality of complex and programmatic assessment approaches.

Methods:

We conducted a concept analysis using Rodgers' evolutionary method to describe the concept of validity as a social imperative in the context of assessment in HPE. Supported by an academic librarian, we developed and executed a search strategy across several databases for literature published between 1995 and 2016. From a total of 321 citations, we identified 67 articles that met our inclusion criteria. Two team members analysed the texts using a specified approach to qualitative data analysis. Consensus was achieved through full team discussions.

Results:

Attributes that characterise the concept were: (i) demonstration of the use of evidence considered credible by society to document the quality of assessment; (ii) validation embedded through the assessment process and score interpretation; (iii) documented validity evidence supporting the interpretation of the combination of assessment findings, and (iv) demonstration of a justified use of a variety of evidence (quantitative and qualitative) to document the quality of assessment strategies.

Conclusion:

The emerging concept of validity as a social imperative highlights some areas of focus in traditional validation frameworks, whereas some characteristics appear unique to HPE and move beyond traditional frameworks. The study reflects the importance of embedding

consideration for society and societal concerns throughout the assessment and validation process, and may represent a potential lens through which to examine the quality of complex and programmatic assessment approaches.

Keywords:

Validity, assessment, health professions education, concept analysis

INTRODUCTION

There have been several calls to reconsider and reshape the role of assessment within health professions education (HPE). Current trends in assessment call for a shift towards assessment *for* learning away from the traditional assessment *of* learning (van der Vleuten et al., 2005), urge increased support for and use of alternative and authentic assessment strategies, and echo a growing call for purposeful and appropriate sampling of performance, all of which are approaches that translate to programmatic assessment (Norcini et al., 2011). One consideration that remains paramount to assessment is that the assessment strategies put in place need to be of the highest quality. The quality of assessment has traditionally been associated with its validity, which is documented through rigorous validation processes that aim to accumulate evidence to support a valid interpretation of the assessment scores (AERA, APA & NCME, 1999; Messick, 1995; Sireci & Sukin, 2013).

The use of traditional approaches to validation for current trends in assessment, such as authentic and programmatic assessment, is increasingly questioned (Berendonk et al., 2013; Gingerich, Kogan, Yeates, Govaerts, & Holmboe, 2014; Ginsburg, McIlroy, Oulanova, Eva, & Regehr, 2010; St-Onge et al., 2017). For instance, there are discussions in the literature suggesting that rater variance may reflect expert judgment and therefore should not be considered solely as measurement error. These discussions suggest traditional measurement models cannot be applied as they consider all rater variance to be measurement error (Lockyer et al., 2017). Additionally, there appears to be some discomfort about relying solely on statistical and psychometric evidence in a validation process for complex assessment such as authentic or programmatic assessment, not because this evidence is inappropriate, but because it may give an incomplete evaluation of the quality of the assessments (e.g., Schuwirth and van der Vleuten (2006), or Hodges (2013)).

In addition to the debates about the applicability of traditional psychometric models to complex assessment, there seems to be little agreement in the HPE community regarding validity itself. St-Onge et al. (2017) relied on discourse analysis methodology to investigate

how validity is used in the HPE literature. The authors identified and describe three different conceptualisations of validity within the HPE literature: (i) *validity as a test characteristic* (similar to the classic validity ‘trinity’ [content, criteria, and construct]); (ii) *validity as an argument-based evidentiary chain* (similar to the argument-based approaches favoured in Messick’s (1995) and Kane’s frameworks (2006)), and (iii) *validity as a social imperative* (St-Onge et al., 2017). By contrast with the other conceptualisations, *validity as a social imperative* is an emerging discourse in the HPE literature and reflects a more holistic vision of assessment and validation, specifically highlighting considerations for broader individual (learner) and societal consequences of assessment. For example, within the discourse analysis, consideration for ‘societal consequences’ included the necessity of assuring the quality of high-stakes assessments leading to the certification of health professionals.

The emerging concept of validity as a social imperative was represented across a body of literature and reflected discussions around programmatic assessment and increased consideration of the potential consequences of assessment for different stakeholders in society (e.g. learners, assessors, educational programs, the broader public). This discourse is the least concretised or operationalised in St-Onge et al.’s (2017), probably because of its emergent nature. Hence, validity as a social imperative requires a deep examination in order that its breadth and scope within the HPE literature can be more fully vetted. Here, we apply Rodgers’ approach to concept analysis (Rodgers & Knafl, 2000; Tofthagen & Fagerstrøm, 2010) in order to clarify the emerging concept of validity as a social imperative.

METHODS

Concept analysis is a methodology used in the development of knowledge (Rodgers & Knafl, 2000; Tofthagen & Fagerstrøm, 2010). Through a structured and rigorous process, this methodology can be used to clarify and refine a concept so that it can be used more efficiently in practice (Rodgers & Knafl, 2000). The method developed by Rodgers (1989) is inductive and relies on qualitative approaches to data analysis (Rodgers, 1989; Rodgers & Knafl, 2000). Rodgers illustrates her method as a spiral to show the constant evolution

of a concept through an iterative process (Rodgers, 1989; Rodgers & Knafl, 2000), and this constant evolution resonated well with our purpose {Formatting Citation}.

Rodgers describes six steps in completing an evolutionary concept analysis, which can be performed in a non-linear fashion (Rodgers & Knafl, 2000). Table 1 briefly describes these steps and their alignment with more traditional research terminology. Whereas Step 1-4 align most directly with the methods section, Step 6 is described in the conclusion of this article. Step 5 is not always included for evolving concepts and, given the state of progress of our work, we have preferred to include it as an avenue for future work.

Table 1
Steps in Rodgers' method (Tofthagen & Fagerstrøm, 2010; Wills & McEwen, 2011)

Steps	Rodgers' concept analysis steps	Parallel research terminology
1	Identify the concept of interest and associated expressions	Research question or topic
2	Identify and select an appropriate realm for data collection	Setting and sampling
3	Collect data relevant to identifying attributes, antecedents and consequents of the concept	Collection and management of the data
4	Analyse data for the attributes, antecedents and consequents of the concept	Data analysis
5	Identify an example of the concept, if appropriate	Demonstration of the concept
6	Identify implications and hypotheses for further development of the concept	Implications of research

Step 1- Identify the concept of interest and associated expressions

It is essential to appropriately select the terminology (word or group of words) used to label the concept to be defined (Rodgers & Knafl, 2000). Initially, the emerging concept of validity as a social imperative (St-Onge et al., 2017) was developed from HPE literature. In the execution of this concept analysis, we found it important to specify the context; therefore, the concept that is the core of this study is validity as a social imperative for assessment in HPE.

Step 2- Identify and select an appropriate realm for data collection

Drawing from the original terms used by St-Onge et al. (2017), and with the assistance of an academic librarian, a search strategy was developed to target articles that included each of the following concepts: (i) validity; (ii) social imperative; (iii) assessment, and (iv) health professional (e.g. medicine, nursing or physical therapy professional). These concepts were operationalised to a search strategy relying on various combinations of keywords such as ‘valid*’, ‘learning outcome*’, ‘social consequence*’, ‘consequential validity’, ‘inference*’, ‘assess*’ and ‘health profession*’. The full list of search terms is available online in Appendix S1. This search was performed in relevant databases, including Academic Complete search, ERIC (Educational Resources Information Center), Education Source, CINAHL (Cumulative Index to Nursing and Allied Health Literature), Francis, Pascal, PsychINFO, and MEDLINE.

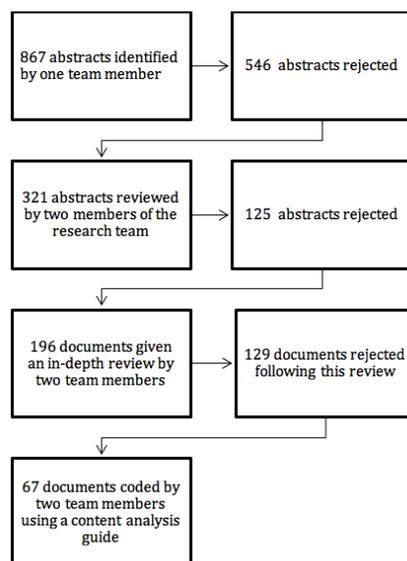
We limited articles to those published during the period from 1995 to 2016 inclusive. The concept of validity as a social imperative is akin to Messick’s idea of evidence of consequences (1995) resulting from the interpretations of assessment scores; therefore his work anchored the 1995 cut-off in our search. Different types of documents in English or French were included in the review, such as research papers, scientific articles, reference books, editorials or letters of opinion, and systematic reviews. To be included in the study, documents were required to discuss validity, or the validation of assessment strategies or programmatic assessment of learners, specifically with a focus on the social responsibility associated with ensuring the quality of assessment or consequences of assessment in the context of HPE. This list of inclusion criteria was based on the initial description and identification of features of the discourse of validity as a social imperative by St-Onge and colleagues (2017). Documents excluded were, for example, about the validity of clinical tools, the quality of learning strategies, and programme evaluation or needs assessments. Specific inclusion and exclusion criteria were developed and iteratively reviewed by the research team (Appendix S2).

Figure 1 illustrates the document selection procedure. The application of the search strategy in the selected databases identified 867 articles, the titles and abstracts of which were

reviewed. A first reading of the titles and abstracts resulted in the exclusion of 546 documents. Documents removed were either duplicate records or articles that did not meet the aforementioned inclusion criteria or were not accessible (i.e. we were unable to obtain paper or electronic versions).

After this initial selection, a second review of the remaining 321 abstracts was conducted by two team members. This second reading helped us to further refine the inclusion and exclusion criteria and to remove documents that did not meet the final inclusion criteria (e.g. items that were not about validity of assessment [n=40] or not about HPE [n=12]). In the event of disagreement between the two reviewers, inclusion was determined by discussion followed by consensus with the entire research team. This resulted in the exclusion of 125 abstracts. A first reading of the full text of 196 selected documents was conducted by two members of the research team to ensure the relevance of the documents to this study. A total of 67 documents were included in this study for in-depth analysis.

Figure 1
Document selection process



Step 3 - Collect data relevant to identifying attributes, antecedents and consequents of the concept

We present steps 3 and 4 separately; however, it is important to note that during the concept analysis, data collection and management are performed as an iterative process (Rodgers & Knafl, 2000). Data collection and analysis required several iterative rounds of discussion among team members to clarify the concept.

All 67 documents selected were read by two members of the team to identify data relevant to the concept's attributes, antecedents and consequents. Any component of an included document could be considered as a source of data: the abstract, introduction, methods, results, discussion and conclusion could all generate data relevant for the refinement of the concept of validity as a social imperative. In view of the emergent status of the concept, we preferred to retain all texts that met our inclusion criteria and did not engage in an analysis of the quality of individual texts. The identification of attributes, antecedents, and consequents is a key feature of a concept analysis, as described by Rodgers (1989). These main areas of focus helped us to collect, organise and manage the data before the analysis. More concretely, *antecedents* are what precedes the concept (Rodgers & Knafl, 2000) of validity as a social imperative for assessment in HPE. *Attributes* are the characteristics that define the concept (Rodgers & Knafl, 2000) or what distinguishes or characterises validity as a social imperative for assessment in HPE. The *consequents* of a concept are what happens as a direct result of the concept (Rodgers & Knafl, 2000).

Step 4- Analyse data regarding the attributes, antecedents and consequents of the concept

Although concept analysis was the overarching methodological frame, it is not prescriptive regarding which qualitative analysis approach should be used to interpret data gathered from texts. Here, we decided to use the qualitative analysis process described by Miles et al. (2014) because it allows for the identification of emerging codes (inductive process) within predetermined codes, which, given the structure of a concept analysis, includes antecedents, attributes and consequents as suggested by Rodgers (Rodgers & Knafl, 2000; Tofthagen & Fagerstrøm, 2010). More specifically, this process consists of three

concurrent analysis cycles performed iteratively and concomitantly: data condensation; data display, and drawing and verifying conclusions.

Data condensation consists of data analysis using predetermined (i.e. antecedents, attributes and consequents) and emerging codes (Miles et al., 2014) (e.g. controversy about consequences of validity or concern about consequences resulting from the interpretation of scores). Two members of the team conducted this analysis cycle. Codes evolved throughout the analysis, but were focused on the overarching targets of identifying antecedents, attributes and consequents of the concept of validity as a social imperative. *Data display* allowed us to better understand and represent our results through data visualisation (Miles et al., 2014). Using a schema, the research team obtained a visual representation of the codes. For example, codes were combine and organised to better align with the structure and focus of a concept analysis (which codes lead to the concept, which codes better describe the concept and which codes may occur from the use of the concept). This overview of all elements helped us to determine whether an excerpt of a document was an attribute, an antecedent, or a consequent. *Drawing and verifying conclusions* relates to the development and verification of conclusions that emerge from data condensation and data display (Miles et al., 2014). Because analysis was an iterative process, the team re-read the documents several times to improve the match between data and codes. Co-analysis and team discussion continued until the team members were satisfied with the interpretation, and the resulting data visualization and summary. The documents were imported into NVIVO FOR MAC VERSION 10.0 NVivo qualitative data analysis Software; QSR International Pty Ltd. Version 10, 2014, to help with the organisation and analysis of the data.

RESULTS

A total of 67 documents were selected and analyzed. Our sample comprised various types of document and predominantly included review articles (n=33, 49%) and research articles (n=27, 40%) (Table 2). Our database included articles from a variety of different disciplines, such as medicine, nursing and physical therapy (Table 3). However, documents reporting on medical education represented more than 50% of our sample.

Table 2
Categories of documents included in the analysis

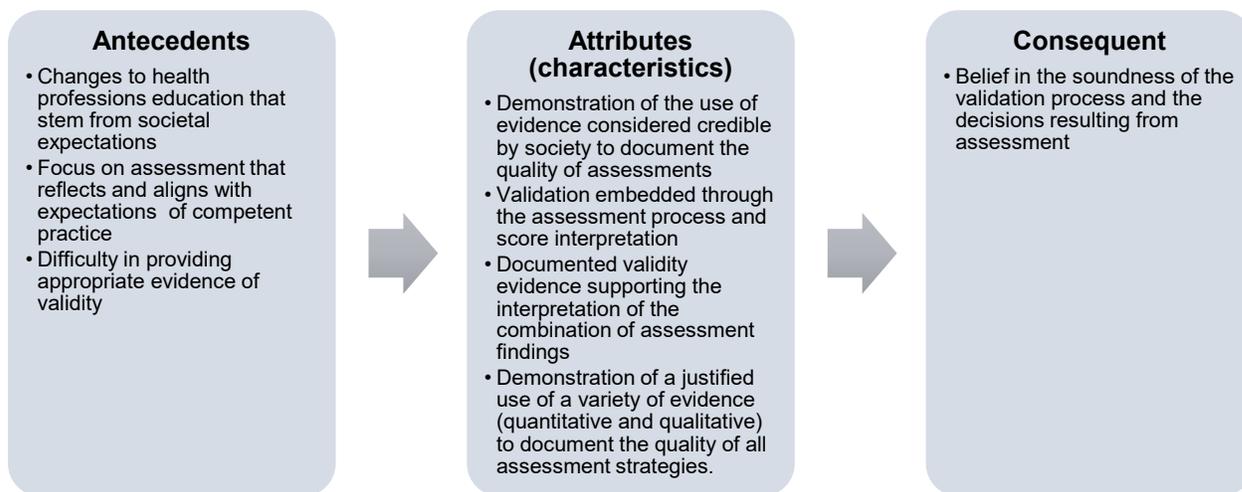
Type of document	n (%)
Review article	33 (49.3%)
Research paper	27 (40.3%)
Editorial, opinion or commentary	4 (6.0%)
Guideline	2 (3.0%)
Literature review	1 (1.5%)
Total	67

Table 3
Disciplines represented in included documents

Discipline	n (%)
Medicine	36 (53.7%)
Nursing	13 (19.4%)
Education	9 (13.4%)
Health Profession	5 (7.5%)
Pharmacy	2 (3.0%)
Social Work	1 (1.5%)
Physical Therapy	1 (1.5%)
Total	67

An overview of the identified antecedents, attributes, and consequents of the concept of validity as a social imperative for assessment in HPE can be found in Figure 2.

Figure 2
Results of the concept analysis of validity as a social imperative
for assessment in health professional education



Here, we will report and describe our ‘main codes’ (antecedents, attributes, consequents), enriched with sample text extracts (verbatim) from the literature included in our concept analysis.

Antecedents

Antecedents are what precedes or has led to the emergence of the concept of validity as a social imperative in the context of assessment in HPE. Based on our analysis, we identified three antecedents for the concept of validity as a social imperative.

Changes to HPE that stem from societal expectations

There has been an increasing desire on the part of society for information on how professional bodies and higher education institutions measure and monitor the acquisition of competencies by health care professionals (Dijkstra, van der Vleuten, & Schuwirth, 2010). For example, one of the core tenets of competency-based education (CBE) is to protect patients by ensuring that professionals achieve a certain level of competency (Arcand & Neumann, 2005; Epstein, 2007; Eva et al., 2015) and are better prepared for the realities of the clinical setting (Baartman, Gulikers, & Dijkstra, 2013; Baartman, Bastiaens, Kirschner, & Van, 2007; Curran et al., 2011). Professional bodies and higher education

institutions must ensure that their assessment programmes are carefully planned and designed in order to measure the abilities of the learners across targeted competencies as accurately as possible.

Focus on assessment that reflects and aligns with expectations of competent practice

To facilitate learning and protect the public, it is necessary to adjust assessment practice (Eva et al., 2015). The literature indicates a shift towards assessment *for* learning from the traditional assessment *of* learning (van der Vleuten & Schuwirth, 2005). We also found a focus on the use of alternative and authentic assessment strategies, characterised notably by realism, simulations of the work context, or mobilisation of the work context itself, and a focus on judgment, complex tasks, and feedback (Palm, 2008; Prigent et al., 2009; Wiggins, 1989). There is a growing use of the term ‘programmatic assessment’ (Norcini et al., 2011), which refers to the combining of assessment strategies, different contexts and several measurement times (van der Vleuten et al., 2012). An assessment programme is constructed deliberately in order to collect sufficient ‘snapshots’ of the future health care professional’s performance (Schuwirth & van der Vleuten, 2011).

Difficulty in providing appropriate evidence of validity

Although psychometric approaches are still valuable and relevant in HPE, traditional validation strategies have limits when it comes to applying them to more modern assessment practices (Govaerts & van der Vleuten, 2013; Schuwirth & van der Vleuten, 2006). The measurement and statistical analysis models currently available to us are appropriate for documenting the quality of individual competence assessment tools, but are not easily extended to the complexities of validation a programme of assessment. Whereas our current psychometric models are well situated for the validation of scores generated by tools, we have few approaches that are appropriate for narrative feedback, or more qualitative assessment data that are increasingly present in HPE (van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). In addition, few psychometric approaches are, indeed, amenable to the validation of entire assessment programmes or of learner’s overall set of competencies (Dijkstra et al., 2010).

The antecedents identified for validity as a social imperative seem to align with ongoing major changes to the training and assessment of health care professionals, most notably the adoption of CBE approaches to training and assessment. We must ensure that the validation approaches used with complex and novel assessment strategies are adapted to document the appropriateness of score interpretation and the quality of the assessment process.

Attributes

Four attributes at the core of the concept of validity as a social imperative (Rodgers & Knafl, 2000) were identified: (i) demonstration of the use of evidence considered credible by society to document the quality of assessments; (ii) validation embedded through the assessment process and score interpretation; (iii) documented validity evidence supporting the interpretation of the combination of assessment findings, and (iv) demonstration of a justified use of a variety of evidence (quantitative and qualitative) to document the quality of all assessment strategies.

Demonstration of the use of evidence considered credible by society to document the quality of assessments

Interpretation of the results of one or more assessment strategies is likely to lead to a decision by the administrators of a given educational programme (e.g. success versus failure). In the context of HPE, poor measurement or misinterpretation of assessment results can have a significant impact on patient safety: a learner who is deemed suitable for practice without the requisite skills or competencies may deliver suboptimal patient care (Cook & Lineberry, 2016; Eva et al., 2015). A learner's success or failure must be based on persuasive assessment data to decrease the potential negative impacts of an ill-prepared trainee on society.

‘This is worrying in the light of evidence that students who are on the borderline in pass/fail decision making may graduate to become doctors whose clinical performance will give cause for concern (Roberts, Newble, Jolly, Reed, & Hampton, 2006, p. 536).’

In other words, high-stakes assessments (such as licensure examinations) may grant the right of practice to trainees who have not yet mastered the competencies required to provide safe care to patients. With such consequences, the professional bodies and higher education institutions responsible for making judgments about a learner's abilities must document the learners' level of competence with sufficient and credible evidence (Beckman, Cook, & Mandrekar, 2005; Colliver, Conlee, & Verhulst, 2012; Cook, Brydges, Ginsburg, & Hatala, 2015; DeLuca, 2011).

Increasingly, professional bodies are transferring responsibility for assessing learner competency to higher education institutions and professional training programs (Baartman et al., 2013). Therefore, establishing validation processes and documenting evidence supporting the quality of assessment is a major aspect of responsibility for higher education institutions, including for program accreditation. Given that the assessment is often used as a guard with which to protect society, we posit that the evidence to document the valid interpretation of assessment scores should be made to society and credible to its users; this takes an external view of the assessment process.

‘A description of how an institution has developed and introduced a defensible high-quality assessment in practice should be of interest to all in medical schools with the responsibility of certifying the competence of their graduates (Roberts et al., 2006).’

Higher education institutions and professional training programmes have a responsibility to society to ensure that they graduate or certify only those trainees who have demonstrated that they are competent practitioners, and must be able to provide evidence to support those decisions.

Validation embedded through the assessment process and score interpretation

Validation is a continuous dynamic process (Schuwirth & van der Vleuten, 2012). Those involved in assessment, which includes development and monitoring, must show constant vigilance for the quality (validity) of their assessments during the development of the assessment and assessment programme, the administration of assessment strategies, and the interpretation of results (Cook et al., 2015).

During the construction of an assessment programme, the deliberate selection of all components of the validation process makes the interpretation of results more credible and more defensible (Berendonk et al., 2013). Building in validity (Ebel, 1983; van der Vleuten et al., 2010) throughout the assessment process helps to improve the quality of all aspects of the assessment programme. In concrete terms, this means that the person responsible for the assessment pre-selects the validation strategies (e.g. the psychometric analyses, interviews with raters), the assessment administration procedure and the training required by the assessors.

‘With overall quality in mind, we advocated that an assessment programme should be constructed deliberately, its elements should be accounted for, it should be centrally governed in its implementation and execution, and it should be regularly evaluated and adapted (van der Vleuten et al., 2012, p. 205).’

Consideration for consequential validity is an essential part of the validation process (Cook & Lineberry, 2016), and should be an important consideration in an effort to build in validity throughout the assessment development and monitoring process. Consequences can be intended or unintended, beneficial or harmful (Kane, 2013; Kane, 2006; Messick, 1995; Sireci & Sukin, 2013). As a result, administrators should consider, and also monitor for, potential consequences of assessment during the development of all assessments included in an assessment programme. More specifically, the consideration of anticipated consequences should include all possible consequences arising from the administration and interpretation of the results, (Cook & Lineberry, 2016) and not be limited in focus by the consideration of only consequences due to construct-irrelevant variance, as typically reported in more classical approaches to validation. Intentional benefits are often the easiest to anticipate and measure, whereas unintentional harms are often the hardest to consider and monitor (Cook & Lineberry, 2016). Consideration for, and identification of, possible negative consequences should allow the assessor, programme and institution to put in place measures or strategies to mitigate them (Hublely & Zumbo, 2011).

Documented validity evidence supporting the interpretation of the combination of assessment findings

In order to support strong assessment practices, validation processes should guide and support the selection of relevant concepts to be assessed, the development of specific tools, the analyses used in a validation approach, and the interpretation of results (AERA et al., 1999). The validation process is also important in a programmatic assessment approach that produces a considerable amount of information in the form of several ‘snapshots’ of the learner collected from different sources collected at different times (Dijkstra et al., 2010). In programmatic assessment, validity evidence must be collected to support the interpretation of the score(s) generated by the assessment programme – evidence supporting the final judgment made based on the assessment data generated within an assessment programme - and not fragmented by individual tools or assessment episodes. This interpretation of the assessment results as a whole is likely to require the judgment of one or more individuals in order to organise and interpret available assessment data.

‘If we cannot break down the results into little pieces that arithmetically add up to a total score, we need to use human judgment to collect and collate information, especially if—in a programme of assessment—information from various types of assessment needs to be combined (Schuwirth & van der Vleuten, 2011, p. 481).’

Validity evidence needs to be documented in order to help assessment administrators make a more defensible decision about the learner’s performance, as represented across the programme of assessment (Dijkstra et al., 2010).

Demonstration of a justified use of a variety of evidence (quantitative and qualitative) to document the quality of all assessment strategies

The quality of the assessment programme must be based on evidence of validity from several sources (Beckman et al., 2005; Colliver et al., 2012; DeLuca, 2011). Evidence must be carefully chosen to ensure that a comprehensive validation process is put in place for any assessment programme (Baartman et al., 2013; Schuwirth & van der Vleuten, 2012). The combination of quantitative and qualitative evidence (mix of methods) may represent a potential avenue of development (Baartman et al., 2013; Schuwirth & van der Vleuten, 2006) for analysis of validity evidence.

‘Such versatility will be required as we move into a ‘post-psychometric era’ of assessment in which qualitative and subjective data are increasingly valued and multiple assessment data points of varying rigour are routinely integrated (Brydges, Ginsburg, Hatala, et al., 2015, p. 561).’

Validity evidence, be it quantitative or qualitative in nature, must be based on a rigorous approach (Cook & Lineberry, 2016; van der Vleuten et al., 2010). To this end, those responsible for the monitoring of programmes of assessment could integrate the corpus of knowledge and methods developed in the context of qualitative research into their validation practices.

‘As in quantitative research, rigour is built into qualitative research, but the terminology and procedures are different. Rigour depends on “trustworthiness” strategies [...] (van der Vleuten et al., 2010, p. 713).’

As an example, administrators responsible for assessment could use different strategies to accumulate information about the learner’s performance in order to triangulate data to look for consistency across different types of evidence or attempt to reach saturation (Cook, Brydges, Ginsburg, & Hatala, 2015; van der Vleuten & Schuwirth, 2005). We could also consider including detailed qualitative data from the individuals who are making ‘overall’ judgments in order to improve the rigor associated with assessment decision-making processes (Hodges, 2013).

These four attributes ([i] demonstration of the use of evidence considered credible by society to document the quality of assessments; [ii] validation embedded through the assessment process and score interpretation; [iii] documented validity evidence supporting the interpretation of the combination of assessment findings, and [iv] demonstration of a justified use of a variety of evidence [quantitative and qualitative] to document the quality of all assessment strategies) are the characteristics that enable us to differentiate the concept of validity as a social imperative in the context of the assessment of learning from other conceptualizations of validity. These attributes contribute to the following consequent.

Consequent

One consequent was identified by the concept analysis: belief in the soundness of the validation process and the decisions resulting from the assessment or assessment programme.

Belief in the soundness of the validation process and the decisions resulting from the assessment

As validity as a social imperative is an emerging concept, there is little literature directly reporting on the topic of its consequents. However, we believe that the consideration for society in the validation processes conducted by different groups (administrators for professional bodies and higher education institutions, as well as the assessors conducting the assessment) will increase trust in the assessment system, in the assessment scores and in judgments resulting from this system. Programmatic assessment needs to be constructed and monitored with relevant, varied, and high-quality evidence of validity (van der Vleuten et al., 2010).

‘Engagement in discussion about performance values by communities of practice may further fuel the debate about what constitutes excellence in professional competence and how assessment systems may contribute to improving the quality of patient care (Govaerts & van der Vleuten, 2013, p. 1172).’

The programme of assessment and validation processes should be planned and explicitly documented; therefore, decisions arising from the interpretation of the assessment results will be more easily defensible and made transparent to all stakeholders, including society at large. Through better selection and documentation of evidence of validity and constant vigilance in order to minimise the consequences arising from the interpretation of the assessment results, those responsible for assessment programmes will be better positioned to support their decisions about learner performance, competence and ability.

DISCUSSION

The purpose of the present study was to clarify the emerging concept of validity as a social imperative in the context of assessment in HPE. Using Rodgers’ concept analysis method (Rodgers & Knafl, 2000; Tofthagen & Fagerstrøm, 2010), we found that attributes

(characteristics) of this concept were: (i) demonstration of the use of evidence considered credible by society to document the quality of assessments; (ii) validation embedded through the assessment process and score interpretation; (iii) documented validity evidence supporting the interpretation of the combination of assessment findings, and (iv) demonstration of a justified use of a variety of evidence (quantitative and qualitative) to document the quality of all assessment strategies.

Our findings are based on an analysis of published literature. Validity as a social imperative may appear to be an outgrowth of contemporary theories of validity to some readers. However, we found reflections, considerations and questions in the literature that go beyond the well-known frameworks and theories. Several elements present in the HPE literature extend beyond traditional validity frameworks: (i) society's concern about the validity of assessment; (ii) the anticipation of consequences of assessment decisions for individuals and society, and (iii) the increased importance of qualitative approaches to validation.

Society's concern about the validity of assessment

Society expects assessments to be of high quality; this applies to assessments ranging from those by higher education institutions (which have the social mandate to train the next generation of health care providers) to those of the various professional bodies (which have the social mandate to protect the public). Several institutions have initiated work to meet this social responsibility and to enable them to respond to societal concerns about assessment. As one example, the Medical Council of Canada (MCC) invested in a process to identify needs relevant to its societal context as relevant to its licensure mandate (i.e. medical practice in Canada). The MCC has recently adapted its assessment blueprint to take into account results from consultations with relevant stakeholders (experts, nurses, pharmacists, physicians and members of the public) (Gotzmann, De Champlain, Touchie, & Boulet, 2013). Validity evidence is necessary to justify decisions resulting from score interpretations, especially in high-stakes assessments, such as admissions, licensing and certification, which represent key decision points for entry into the health care professions.

Eva et al. (2015) highlight that, even with the increased number of high-stakes assessments, the frequencies of medical error and poor quality health care in the USA continue to increase. Societal concerns about patient safety appear to be fuelled by the growing attention given by mass media to events occurring in the health care sector (e.g. adverse events), which could be seen as a reflection of the competency level of health care providers (Carraccio, Englander, Van Melle, et al., 2016). With increasing reliance on programmatic or rater-based assessments, continued recognition of the value of assessment in HPE and an acknowledgement of the high-stakes nature of many assessment decisions, determining the validity of inferences based on assessment scores, including inferences regarding an individual's competence, remains a challenge.

Messick's (1995) unified theory of validity has yet to be operationalised in a way that could help those responsible for the documentation of the quality of their assessment programmes (Bartman et al., 2007; Scoles, Hawkins, & LaDuca, 2003). Although there have been efforts to operationalise Kane's (2013; 2006) argument-based approach to the validation of inferences based on assessment scores for programmatic assessment, this framework does not have the same focus on consequences of assessment as that of Messick (1995). Further work is required to bring the best of these frameworks together in the current context of HPE assessment in order to support validation practices for programmes of assessment.

Anticipation of the consequences of assessment decisions for individuals and society

Assessors must anticipate the potential consequences of assessment decisions on individuals and society, and propose measures to minimise or mitigate them. However, consideration of the potential consequences of the validation process remains controversial. (Downing, 2003; Goodwin, 2002) Some authors such as Mehrens (1997) and Popham (1997), have argued that the consequences of assessment are associated with social policies and not with validity (Goodwin, 2002; Huble & Zumbo, 2011; Kane, 2013; Lane, 2012; Messick, 1989; Sireci & Sukin, 2013). Nonetheless, assessment is a social act and has social consequences. Cook & Lineberry (2016) go so far as to claim that consequential validity is the most important evidence to consider in a validation process. Their

justification is similar to a core feature identified in our analysis: all decisions made based on the results of assessments can impact the individual or society, positively or negatively.

Based on the results of this concept analysis, we believe that consequences should not simply be attended to or measured a posteriori, but considered, attended to and mitigated throughout the validation process (preparation, administration and interpretation of assessment and assessment scores). However, consequences are not frequently measured in HPE (Cook & Lineberry, 2016; Eva et al., 2015; van der Vleuten & Schuwirth, 2005). We found few descriptions of ‘how’ one could measure the consequences of assessment before, during or after the said assessment. Furthermore, measuring consequences is a challenging process that requires substantial effort (Sireci & Sukin, 2013) such as through longitudinal studies to measure the potential unintended consequences of assessment (Eva et al., 2015). Given the difficulty in assuring consequential validity, further studies on the operationalisation of consequential validity are required.

The increased importance of qualitative validation strategies

The shift in HPE towards CBE brought about changes in assessment practice, such as the implementation of programmatic assessment and the increased reliance on performance-based assessment. The literature increasingly suggests the difficulty in providing adequate validity evidence for complex assessment such as authentic and programmatic assessment. As indicated by Hodges (2013), psychometric approaches are currently evolving. We need, as a community, to find solutions for the synthesis and quality monitoring of complex assessment strategies, including integrating qualitative approaches for analysis and synthesis into the validation process. Although the authors of existing validity frameworks (e.g. Messick (1995) and Kane (2006)) do not dismiss the use of qualitative data to support the valid interpretation of assessment scores, little is known of what these data might look like and how they might be considered in a validation process (Hubley & Zumbo, 2011; Kane, 2013; Messick, 1995). A potential risk for increasing the reliance on qualitative validity evidence may suggest a limited place for more traditional quantitative methodologies, with a reduced reliance on traditional psychometric approaches. Some authors have mentioned the need to combine qualitative and quantitative evidence (Cook

& Lineberry, 2016; Crossley, 2013; Kane, 2013). However, the proposed strategies that would align with the goal of combining qualitative and quantitative evidence remain poorly defined or operationalised. The balance between different types of evidence may be difficult to establish and may benefit from consideration of mixed-methods approaches to analysis and data integration (Creswell, 2014; Teddlie & Tashakkori, 2009). Regardless of types of evidence selected, individuals involved in assessment must choose and justify the best evidence to ensure the quality of their programmatic assessment. However, a shift in validation approach is unlikely to be easy, especially for those individuals or institutions with few accessible resources. Whatever the type of validity evidence collected – be it qualitative, quantitative, or mixed - there is, and will always be, a need for human judgment to interpret and make decisions based on these evidences (Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2010).

Strengths and limitations

The use of Rodgers' approach (1989) provided a systematic and rigorous method for analysing and synthesising a large and varied body of literature that is still mainly discursive in nature. We also engaged in various steps to ensure scientific rigour during our concept analysis and qualitative data analysis. Co-coding by two team members ensured both the credibility and confirmability of the data (Morse, 2015; Thorne, 2016). We provided a thick and rich description of the research process to establish transferability and dependability (Thorne, 2016).

This study has limitations. Although we have followed a rigorous process of data collection, it is possible that we missed some literature as a result of our use of specific search terms used to identify literature for the concept analysis. Therefore, certain documents that might have enriched our findings may not have been identified. However, given the emerging nature of the concept of interest, we believe that our understanding of validity as a social imperative will continue to evolve with current literature. A large portion of the literature that was analysed specifically discussed medical education. We had reasonable representation of other disciplines (e.g. nursing, occupational therapy and physical therapy), but some health professions are likely to have been under-represented

or absent from our final database. In a concept analysis, the identification of all relevant literature is not a marker of quality; rather, the goal of a concept analysis is to identify a representative body of literature reflecting the breadth of a concept.

Considering that validity as a social imperative is an emerging concept, the consequent remains largely hypothesised. However, we believe that the description of the consequent reflects what is present in the current literature. Finally, there are overlaps between the concept of validity as a social imperative and other modern frameworks or theories of validity. These overlaps are not surprising as modern validity theories and validation frameworks represent the language currently used when discussions of validity in HPE, and validity as a social imperative stems from current discussions on the topic and includes descriptions of the limitations of the current frameworks. As the concept is still emerging, the definition, boundaries and consequents of this concept are likely to evolve, and the work presented here represents our current understanding based on the existing literature (Toftthagen & Fagerstrøm, 2010). A final limitation of this work is that the analysis of the concept of validity as a social imperative has thus far been restricted to a careful analysis of the published literature. Future research would be well situated to explore the resonance and applicability of this concept with health professions researchers and educators, and work towards its operationalisation for future use.

CONCLUSION

We have gathered and described the attributes of the emerging concept of validity as a social imperative, relying on Rodgers' approach to concept analysis (Rodgers, 1989). Findings from this concept analysis highlight society's growing concern about the quality of assessment and a need for those responsible for assessment to establish the quality of newly emerging assessment approaches. Now that we have clarified the concept, further research can focus on elucidating the intersections between validation approaches, societal concerns, and the operationalisation of this concept according to individuals involved in the development and monitoring of assessment, such as students, assessors, administrators, members of assessment committees, or professional bodies. In light of our results, we posit that there is a need to support administrators through the provision of practical validation

strategies, whether quantitative or qualitative. Findings from our study could be combined or used to complement modern validity theories and as such be used to inform future validation studies in the competency-based new era of HPE. Operationalisation of the concept could better support administrators throughout the various steps in the validation process by providing more practice-oriented validation strategies that align with the concept of validity as a social imperative for assessment in HPE.

Contributors: MM contributed to the design of the study and was primarily responsible for data collection and analysis and the initial drafting of the manuscript. FG, MY and CS-O contributed to the design of the study, and to data collection and analysis. All authors contributed to the critical revision of the paper and approved the final manuscript for publication.

Acknowledgements: The authors would like to thank Marianne Xhignesse, Department of Family Medicine, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Quebec, Canada for her critical review of earlier drafts of this manuscript.

Funding: This study was supported by the Chaire de Recherche en Pédagogie Médicale Paul Grand'Maison de la Société des Médecins de l'Université de Sherbrooke, the MEES-Université programme, and the Social Sciences and Humanities Research Council (in a grant no. 435-2014-2159 provided to CS-O and MY).

Conflicts of interest: None.

Ethical approval: Not applicable

References

- AERA, APA, & NCME. (1999). Validity. In *Standards for Educational and Psychological Testing* (pp. 9–24). Washington, DC.
- Arcand, L. L., & Neumann, J. A. (2005). Nursing Competency Assessment Across the Continuum of Care. *Journal of Continuing Education in Nursing*, *36*, 247–254.
- Baartman, L., Gulikers, J., & Dijkstra, A. (2013). Factors influencing assessment quality in higher vocational education. *Assessment & Evaluation in Higher Education*, *38*(8), 978–997. <https://doi.org/10.1080/02602938.2013.771133>
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C.P.M. (2007). Teachers' opinions on quality criteria for competency assessment programs. *Teaching and Teacher Education: An International Journal of Research and Studies*, *23*(6), 857–867.
- Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal Of General Internal Medicine*, *20*(12), 1159–1164.
- Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. T. (2013). Expertise in performance assessment: Assessors' perspectives. *Advances in Health Sciences Education*, *18*(4), 559–571. <https://doi.org/10.1007/s10459-012-9392-x>
- Carraccio, C., Englander, R., Van Melle, E., ten Cate, O., Lockyer, J., Chan, M.-K., ... Snell, L. S. (2016). Advancing Competency-Based Medical Education. *Academic Medicine*, *91*(5), 645–649. <https://doi.org/10.1097/ACM.0000000000001048>
- Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity ... and back? *Medical Education*, *46*(4), 366–371.
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: a practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575.
- Cook, D. A., & Lineberry, M. (2016). Consequences validity evidence: evaluating the impact of educational assessments. *Academic Medicine*, *91*(6), 785–795.
- Creswell, J. W. (2014). *Research Design : Qualitative, Quantitative, and Mixed Methods Approaches*. SAGE Publications.
- Crossley, J. (2013). Validity and truth in assessment. *Medical Education*, *47*(12), 1152–1154.
- Curran, V., Hollett, A., Casimiro, L. M., Mccarthy, P., Banfield, V., Hall, P., ... Wagner, S. (2011). Development and validation of the interprofessional collaborator assessment rubric ((ICAR)). *Journal of Interprofessional Care*, *25*(5), 339–344.
- DeLuca, C. (2011). Interpretive validity theory: mapping a methodology for validating educational assessments. *Educational Research*, *53*(3), 303–320.
- Dijkstra, J., van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education*, *15*(3), 379–393.
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837.
- Ebel, R. L. (1983). The practical validation of tests of ability. *Educational Measurement: Issues and Practice*, *2*(2), 7–10.

- Epstein, R. M. (2007). Medical education - assessment in medical education. *New England Journal of Medicine*, 356(4), 387–396.
- Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E., & Regehr, G. (2015). Towards a program of assessment for health professionals: from training into practice. *Advances in Health Sciences Education*, 21(4), 1–17.
- Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the “black box” differently: assessor cognition from three research perspectives. *Medical Education*, 48(11), 1055–1068.
- Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward authentic clinical evaluation: pitfalls in the pursuit of competency. *Academic Medicine*, 85(5), 780–786.
- Goodwin, L. D. (2002). Changing conceptions of measurement validity: an update on the new standards. *The Journal of Nursing Education*, 41(3), 100–106.
- Gotzmann, A., De Champlain, A., Touchie, C., & Boulet, J. (2013). *National Survey of the Physicians, Pharmacists, Nurses, and Public in Canada: 2013*. Ottawa: Medical Council of Canada.
- Govaerts, M. J. B., & van der Vleuten, C. P. M. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education*, 47(12), 1164–1174.
- Hodges, B. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*, 35(7), 564–568.
- Hubley, A. M., & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research*, 103(2), 219–230.
- Hupcey, J. E., Morse, J. M., Lenz, E. R., & Tasón, M. C. (1996). Wilsonian methods of concept analysis: a critique. *Scholarly Inquiry for Nursing Practice: An International Journal*, 10(3), 185–210.
- Kane, M. (2013). The argument-based approach to validation. *School Psychology Review*, 42(4), 448–457.
- Kane, M. T. (2006). Validation. In: Brennan RL, ed. *Educational Measurement*. 4th edn. Westport, CT: American Council on Education/Praeger; 17–64.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73.
- Lane, S. (2012). Consequences of assessment and accountability systems are integral to the argument-based approach to validity. *Measurement: Interdisciplinary Research & Perspective*, 10(1–2), 71–74.
- Lockyer, J., Carraccio, C., Chan, M.-K. K., Hart, D., Smee, S., Touchie, C., ... Frank, J. R. (2017). Core principles of assessment in competency-based medical education. *Medical Teacher*, 39(6), 609–616.
- Mehrens, W. A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice*, 21(1), 16–18.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan.
- Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8.
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis*. Sage.
- Morse, J. M. (2015). Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qualitative Health Research*, 25(9), 1212–1222.

- Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206–214.
- Palm, T. (2008). Performance Assessment and Authentic Assessment: A conceptual Analysis of the Literature. *Practical Assessment Research Evaluation*, 13(4), 1–24.
- Popham, W. (1997). Consequential validity: right concern-wrong concept. *Educational Measurement: Issues and Practice*, 21(1), 9–13.
- Prégent, R., Bernard, H., & Kozanitis, A. (2009). *Enseigner à l'université dans une approche-programme*. Montreal, QC: Presses Internationales Polytechnique.
- Roberts, C., Newble, D., Jolly, B., Reed, M., & Hampton, K. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher*, 28(6), 535–543.
- Rodgers, B. L. (1989). Concepts, analysis and the development of nursing knowledge: the evolutionary cycle. *Journal of Advanced Nursing*, 14(4), 330–335.
- Rodgers, B. L., & Knafl, K. A. (2000). *Concept Development in Nursing*. Philadelphia, PA: WB Saunders.
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), 38–48.
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). Programmatic assessment: from assessment of learning to assessment for learning. *Medical Teacher*, 33(6), 478–485.
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, 40(4), 296–300.
- Scoles, P. V., Hawkins, R. E., & LaDuca, A. (2003). Assessment of clinical skills in medical practice. *Journal of Continuing Education in the Health Professions*, 23(3), 182–190.
- Sireci, S. G., & Sukin, T. (2013). Test validity. In: Geisinger KF, ed. *APA Handbook of Testing and Assessment in Psychology Vol 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology*. Washington, DC: American Psychological Association, 61–84.
- St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education*, 22(4), 853–867.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of Mixed Methods Research : Integrating Quantitative and Qualitative Approaches in the Social and Behavioral Sciences*. Thousand Oaks, CA: Sage Publications.
- Thorne, S. (2016). *Interpretative Description: Qualitative Research for Applied Practice*. New York, NY: Routledge.
- Toftagen, R., & Fagerström, L. M. (2010). Rodgers' evolutionary concept analysis - a valid method for developing knowledge in nursing science. *Scandinavian Journal of Caring Sciences*, 24(SUPPL. 1), 21–31.
- van der Vleuten, C.P.M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205–214.
- van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, 39, 309–317.

- van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 24(6), 703–719.
- Wiggins, G. (1989). A True Test: Toward More Authentic and Equitable Assessment. *Phi Delta Kappan*, 70(May), 703–713.
- Wills, E., & McEwen, M. (2011). Concept development clarifying meaning of terms. In: McEwen M, Wills E, eds. *Theoretical Basis for Nursing*, 3rd edn. Philadelphia, PA: Lippincott, Williams and Wilkins 2011; 46–67.

Appendix S1
Search terms

Concept 1		Concept 2		Concept 3		Concept 4
Validity		Social imperative		Assessment		Health professionals
Valid* Or Standard* Or Test* Or Quality*	AND	“learning outcome*” Or “consequential validity” Or “educational achievement” Or “social consequence*” Or “Test interpretation “ Or “test perception” Or “value implication*” Or “student performance” Or “problems of professional competence” Or Inference*	AND	evaluation* Or assessment* Or “educational evaluation*” Or skill* Or competence* Or aptitude* Or “program satisfaction” Or “admission exam” Or “program climate” Or “continuing professional development” Or “continuing education” Or “competency-based education” Or “abilities-based curriculum” Or “competency-by-design”	AND	“higher education” Or “undergraduate student*” Or “health professional*” Or “medical education”

Appendix S2
Inclusion and exclusion criteria

Selection criteria	Inclusion	Exclusion
Year of publication	≥ 1995	
Language	English or French	
Source	Research papers Scientific articles Reference books Editorials or letters of opinions Systematic reviews	Paper or electronic version not available
Level of education	Undergraduate Graduate Postgraduate	Preschool Primary education Junior and senior secondary education
Discipline	Health professions (e.g., Medicine, Nursing, Physical Therapy, Occupational Therapy) Education	Other disciplines such as: Animal Health Administration Communication Literature Law
Purpose of assessment	Learners Teachers Validity of assessment strategies Validity of programmatic assessment Social accountability of assessment	Patients Validity related to clinical interventions, clinical or research tools, program assessment, or needs assessment

5. MÉTHODE, RÉSULTATS ET DISCUSSION DE LA PHASE 2 – ACCEPTABILITÉ ET FAISABILITÉ ANTICIPÉE

Dans le chapitre 5, nous décrivons le deuxième objectif de la recherche qui visait à explorer l'acceptabilité et la faisabilité anticipée du concept de validité en tant qu'impératif social. Tout d'abord, nous présenterons la méthode, les résultats et la discussion sous la forme d'un deuxième article. Le nombre de mots étant limité lors de la rédaction d'un article, nous décrivons par la suite les éléments méthodologiques et leurs justifications que nous n'avons pu expliciter. Nous compléterons également la description des participants par la présentation des caractéristiques sociodémographiques.

5.1 Article 2

Validity as a social imperative: Users' and leaders' perceived acceptability and anticipated feasibility

Auteurs de l'article

Marceau, Mélanie; St-Onge, Christina; Gallagher, Frances; Young, Meredith

Statut de l'article

Soumis le 8 août 2019 à *Perspectives on Medical Education*

Avant-propos de l'article

Nous décrivons dans cet article la méthode, les résultats et la discussion de l'objectif 2 de l'étude doctorale qui vise à explorer l'acceptabilité et la faisabilité anticipée du concept de validité en tant qu'impératif social auprès de différentes parties prenantes impliquées dans le domaine de l'évaluation des apprentissages en pédagogie des sciences de la santé. Nous souhaitons aviser le lecteur que, pour répondre aux exigences de la revue, l'article est rédigé en anglais britannique.

Contribution de l'étudiante et des coauteurs

L'étudiante a participé à la collecte des données. Elle a élaboré les guides d'entrevues et mené les entrevues individuelles et de groupe chez le personnel enseignant et les membres de comités d'évaluation des apprentissages. Pour les personnes engagées en recherche au Canada, les entrevues ont été menées par une professionnelle de recherche. Les données qualitatives ont été coanalysées par l'étudiante et ses directrices. L'étudiante a rédigé le premier jet du manuscrit et a procédé au raffinement du contenu, qui a été bonifié par les réflexions et commentaires de l'ensemble des auteures. Les directrices de recherche ont supervisé l'étudiante pendant la collecte des données et l'analyse des données. Elles ont également contribué à la rédaction de l'article.

Validity as a social imperative: Users' and leaders' perceived acceptability and anticipated feasibility

Marceau, Mélanie, MSc, PhD candidate, St-Onge, Christina. PhD., Gallagher, Frances, PhD. and Young, Meredith, PhD.

M. Marceau is a PhD candidate, Department of Nursing, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada.

C. St-Onge is Professor, Department of Medicine, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada. She holds the *Paul Grand'Maison de la Société des médecins* Research Chair in Medical Education.

F. Gallagher is Professor, Department of Nursing, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada.

M. Young is Associate professor, Department of Medicine and Institute for Health Sciences Education, Faculty of Medicine, McGill University, Montreal, Québec, Canada.

Résumé

Introduction: Plusieurs conceptualisations de la validité sont utilisées en pédagogie des sciences de la santé pour documenter la qualité de l'évaluation. Récemment, la validité en tant qu'impératif social a été proposée comme une conceptualisation émergente de la validité. Comme première étape de la mise en œuvre du concept de validité en tant qu'impératif social, nous avons exploré l'acceptabilité et la faisabilité anticipée du concept auprès des utilisateurs et des leaders impliqués dans l'évaluation des apprentissages en pédagogie des sciences de la santé au Canada.

Méthode: Une étude qualitative descriptive interprétative a été réalisée. Des stratégies d'échantillonnage par choix raisonné et par boule de neige ont servi au recrutement des participants aux entrevues individuelles semi-dirigées et aux groupes de discussion focalisée. Chaque transcription a été analysée par l'étudiante et une directrice de recherche puis discutée avec l'ensemble de l'équipe jusqu'à l'obtention d'un consensus.

Résultats: Trente personnes provenant de deux groupes de parties prenantes (utilisateurs et leaders) ont participé à notre étude. Selon nos résultats, le concept de validité en tant qu'impératif social est généralement perçu comme acceptable. Quel que soit leur groupe, les participants ont partagé des considérations similaires concernant: les limites des processus de validation existants, la pertinence du concept, la nécessité de clarifier certains termes utilisés pour définir le concept, les similitudes avec les théories modernes de la validité et les défis anticipés lors de l'application du concept dans la pratique. En outre, les participants ont discuté de certaines limites des processus de validation existants dans le contexte de l'évaluation programmatique et lors de l'évaluation en milieu de travail.

Discussion: La validité en tant qu'impératif social semble complémentaire aux théories existantes de la validité et peut représenter une façon d'adapter les modèles dits traditionnels de la validité de façon à prendre en compte la complexité de l'évaluation des apprentissages en pédagogie des sciences de la santé. Cependant, des difficultés sont envisagées sur le plan de l'opérationnalisation du concept.

Abstract

Introduction: Several conceptualisations of validity are used in health professions education (HPE) to document the quality of assessment. Recently, *validity as a social imperative* was proposed as an emerging conceptualisation of validity. As a first step to operationalisation, we explored the perceived acceptability and anticipated feasibility of validity as a social imperative with users and leaders engaged with assessment in HPE in Canada.

Methods: A qualitative interpretive description study was conducted. Purposeful and snowball sampling were used to recruit participants for semi-structured individual interviews and focus groups. Each transcript was analysed by two team members and discussed with the team until consensus was reached.

Results: Thirty individuals from two different stakeholder groups (users and leaders) participated in our study. Our findings suggest that the concept of validity as a social imperative is generally perceived as acceptable. Regardless of their group, participants shared similar considerations regarding: the limits of traditional validity models, the concept's timeliness and relevance, the need to clarify some terms used to characterise the concept, the similarities with modern theories of validity, and the anticipated challenges in applying the concept in practice. In addition, participants discussed some limits with current approaches to validity in the context of workplace-based and programmatic assessment.

Discussion: Validity as a social imperative seems complementary to existing theories of validity and, may represent how HPE is adapting traditional models of validity in order to respond to the complexity of assessment in HPE; however, challenges likely remain in operationalising the concept prior to its implementation.

What this paper adds

Current validation practices fall short when documenting the quality of assessment practices such as Entrustable Professional Activities, milestones, and programmatic assessment. The gap between currently available assessment and validation practices contributed to an emerging conceptualisation of validity being documented in the literature - validity as a social imperative. In this conceptualisation, great importance is given to the impact of assessment and score interpretations on individual students and society as a whole, and it appears well suited for new assessment practices. In this paper, we explored the acceptability and perceived feasibility of this new concept from two different key stakeholder groups.

INTRODUCTION

Educators and researchers have witnessed the rise of narrative assessment [1], rater-based and workplace-based assessment [2,3], programmatic assessment [4,5], and the integration of a variety of different data sources in the context of a comprehensive validation process [5,6]. Unfortunately, current validation practices may not be well suited for recent changes in assessment practices, and therefore, new ways of thinking about- and doing- validity are required. Marceau, Gallagher, Young and St-Onge [7] suggested that an emerging conceptualisation of validity, validity as a social imperative, may offer a complimentary avenue to traditional psychometrics analyses for more recent assessment practices such as programmatic and workplace-based assessment. Marceau et al. [7] identified four characteristics of the concept of validity as a social imperative that are briefly described in Table 1.

Table 1
 Characteristics of the concept of validity as a social imperative

Characteristics	Definition
Validity evidence seen as credible by society	Professional bodies (teaching institutions and professional orders) must be able to document, in a way that is perceived as credible by society, decisions made regarding a students' knowledge, attitude, skills and competencies.
Validity built into the assessment process	Validity evidence includes the justification of decisions made during the development and administration of an assessment, and the interpretation of assessment results. This evidence includes consideration of the potential consequences that the interpretation of the assessment scores could have on the individual, the institution, and society.
Interpretation of the combination of assessment findings	Assessment data generated within an assessment programme are often combined to make a final judgment. Validity evidences should be collected to support the combined or total score interpretation– evidences should align with the intended score use.
Validity evidence includes quantitative and qualitative data	Validity evidence must be collected using rigorous approaches, and attention should be paid to quantitative and qualitative data sources as legitimate validity evidence.

Adapted from Marceau et al. [7]

Our current understanding of the concept of validity as a social imperative is based on analyses of published literature [7,8]. We know little regarding if, and how, this emerging conceptualisation of validity is understood, nor how it is perceived by members of the health professions education (HPE) community engaged in conversations about assessment, validity and validation. Understanding the perceived strengths, weaknesses, anticipated practical challenges, and implications of validity as a social imperative for assessment in HPE is critical for further refinement of the concept, and to engage stakeholders to help shape its operationalisation for use. In this study, we explored the perceived acceptability and anticipated feasibility of the concept of validity as a social imperative with users and leaders engaged with assessment in HPE in Canada.

METHODS

Design overview

We conducted a qualitative interpretive description study adapted from Thorne's approach [9–11] in order to ground findings in practice and to generate meaningful results. Semi-structured interviews were carried out with individuals from two stakeholder groups 1) users such as educators and assessment committee members and 2) leaders engaged in HPE assessment and scholarship (subsequently referred to as leaders). These stakeholder groups were selected to reflect the diversity within Canadian assessment communities in HPE. The study was conducted in Canada which has been a leader in research in HPE (e.g.: number of publications per medical school), in assessment and validity in medical education specifically [12], and in medical education more broadly [13]. A brief description of participants, procedure and study timelines for the two stakeholder groups in Table 2.

Table 2
Procedure used with stakeholders

Stakeholder	Users	Leaders
Participants	Educators and assessment committee members	Individuals involved in the Canadian HPE research and assessment community
Sampling	Purposeful sampling	Purposeful sampling Snowball sampling
Recruitment	Email to educational program directors	Direct email
Duration	October and November 2016	January and July 2017
Method	Focus group and individual interviews Length 60-90 minutes	Individual phone interviews by two research professionals Length 60-90 minutes
Material	Sociodemographic questionnaire Semi-structured interview guide	Sociodemographic questionnaire Semi-structured interview guide

Ethical consideration

We obtained ethical approval from the board of Education and Social Sciences, Université de Sherbrooke (2016-34-ESS) and from the other universities where the research was conducted. Consent was obtained prior to the start of the interview; no compensation was

provided. Participation was voluntary, and participants could withdraw at any time. Participants from focus groups were asked to preserve the confidentiality of the content of the discussions. The research team ensured the confidentiality of the data collected and anonymised the transcripts.

Participants and Recruitment

Users

We recruited educators and committee members using purposeful sampling to encourage different perspectives (four different programs from four universities). To be included, participants had to be professors, lecturers or assessment committees' members for more than one year. They had to be involved in nursing, medicine, physiotherapy or occupational therapy programs the targeted Québec Universities to facilitate face-to-face focus groups. We included participants engaged in the design, validation, or monitoring of assessment strategies within their local context. We excluded individuals who deliver a few lectures, tutors, instructors or those who solely supervise health professions trainees.

Leaders

We purposefully approached individuals involved in the Canadian HPE research and assessment community. To build our list of potential participants, we included researchers recognised for their contributions to assessment or validity and individual with governance roles in assessment in HPE. We purposefully considered individuals across the professional spectrum (e.g. junior to senior researchers), representing various domains of work in HPE (e.g. licensure programs, membership on assessment policy committees, undergraduate, postgraduate), and geographic distribution. Snowball sampling was used to expand our list of potential participants. Participants in our sample had published an average of 100 publications relevant to HPE, and hold or have held leadership positions within HPE or within their specific research domain. There were no exclusion criteria.

Data collection

Material

Preparatory documents. Prior to the interview, all participants received two documents: 1) a two-page summary of three conceptualisations of validity in HPE, including a description of validity as a social imperative through the lens of a discourse analysis [8]; and 2) a two-page summary of the results from the concept analysis describing the antecedents, characteristics, and consequences of the concept of validity as a social imperative [7]. Participants were expected to review these documents before the interview.

Sociodemographic questionnaire. Participants completed a short sociodemographic questionnaire asking: their age group, gender, and initial training and experience in the field of assessment and validity.

Semi-structured interview guide. The semi-structured interview guide (see appendix 1 for users, and appendix 2 for leaders) was comprised of key open-ended questions, using probes as needed, and adapted as relevant for the different stakeholder groups [14,15]. Participants were asked to describe their views concerning the acceptability and feasibility to the characteristics of validity as a social imperative (Table 1) and share their opinion on the concept of validity as a social imperative, and any further comments.

Procedure

Users

We emailed educational program directors and asked them to circulate a short description of the study and an invitation to participate to relevant individuals. Interested individuals contacted us to obtain a copy of the consent form and preparatory documents. MM collected data through in-person focus groups and semi-structured individual interviews. We aimed to use focus groups to gather different perspectives enriched by the interaction between participants [16], but due to participant availability, some were interviewed individually. Focus groups and individual interviews lasted 60 to 90 minutes. All interviews were recorded and transcribed.

Leaders

We directly contacted participants by email. Those who expressed interest received additional information by email including a consent form, a sociodemographic questionnaire, and the above-mentioned preparatory documents. Two experienced research professionals (KD, LA) conducted semi-structured individual phone interviews. Interviews lasted 60 to 90 minutes, were recorded and transcribed. Interviewers received coaching (from MM) throughout the study to ensure data collection quality. We used interviewers outside the research team in order for participants to feel as though they could speak critically.

Data Analysis

Descriptive statistics were used to summarise participant characteristics. Results from each stakeholder group were analysed separately. Two team members (MM, FG) carried out the qualitative analysis and all team members discussed the interpretation until a consensus was reached [14]. Data organisation and management were facilitated by Dedoose version 8.0.31 (SocioCultural Research Consultants, Los Angeles, CA). The analysis was guided by the three concurrent analysis cycles in Miles, Huberman and Saldana's method [17].

Data condensation: Our study was informed by Sidani & Braden's [18] definitions of acceptability (perception of the concept regarding relevance, convenience, effectiveness, and risks associated with the concept or the adherence of the concept) and feasibility (the possible application of the concept into practice). Through inductive process, additional codes were added throughout analysis [14,17].

Data display: We explored participants' perceived acceptability and anticipated feasibility for the four characteristics of validity as a social imperative individually (described in Table 1). Conceptually clustered matrices helped to organise and visualise data to draw conclusion [17] and make visible discrepancies, similarities and relations, whether between the different characteristics of the concept or between the stakeholders.

Drawing and verifying conclusions: Regular team discussion and reflection throughout the analysis process helped to enhance the interpretation of the results. After each leader interview, the research team synthesised main themes in the transcript, identified exemplary quotes, and shared this summary with participants for review. Leaders were given the opportunity to expand, adapt, or suggest modifications to the summary as needed via email [19,20]. One participant requested adjustments to the summary to provide more context supporting an exemplary quote.

Team composition

The research team has been working on the concept of validity as a social imperative in the context of MM's PhD and bring different expertise. FG is an established scholar in nursing, with expertise in the qualitative approaches used. CSTO and MY are established scholars in medical education, with areas of research on assessment and validity in health professions education. Two experienced research assistants (KD, LA) conducted the interviews with leaders.

RESULTS

Participants:

Twenty-three users and seven leaders participated. We conducted five focus groups (n = 19 participants; 3-8 participants per group) and four individual interviews with users. All leaders participated in individual interviews. Participant characteristics are presented in Table 3.

Table 3
Sociodemographic characteristics of participants

Characteristics	Users (n=23)	Leaders (n=7)
Gender n (%)		
Men	12 (52%)	5 (71%)
Women	11 (48%)	2 (29%)
Initial training n (%)		
Medicine	7 (30%)	3 (43%)
Nursing	8 (35%)	-
Physiotherapy	2 (9%)	-
Occupational Therapy	2 (9%)	-
Education	4 (17%)	1 (14%)
Others (e.g, psychology)	4 (17%)	3 (43%)

Findings:

During the interviews, we noticed that educators and assessment committee members tended to move between discussing issues of assessment and issues of validity throughout the interviews – demonstrating the interdependence of the two concepts. To synthesise our findings, we report similarities and differences across participant groups (Table 4).

Table 4
Users and leaders' perspectives

Description	Themes
Similarities across stakeholder groups	Relevance of the concept in the current context
	Required clarification of terms used in the concept analysis
	Resemblance with modern theories of validity
	Challenges in the application of the concept to practice
Differences between the two stakeholder groups	Differing conceptualisations of the importance of assessment and validity
	Society as a driving force for different ends

Similarities across stakeholder groups

Participants shared similar opinions regarding: 1) the relevance of the concept in the current context; 2) the need to clarify some terms used in the concept analysis; 3) the similarities with modern theories of validity and 4) the challenges related to the application of the concept in practice.

Relevance of the concept in the current context

Participants suggested that within the concept of validity as a social imperative there is an opportunity to explore validation practices that do not solely rely on psychometric analyses. Participants highlighted the importance of considering the overall assessment result obtained through programmatic approaches to assessment.

Integrating it [validity] as part of an entire program, that would be the big difference. The modern validity or traditional validity, you can focus on just an exam itself and just a rating scale itself ... as a part (...) thinking about the whole and what the whole means, that's a new thing. (L7)²

Participants suggested that the uptake or consideration for validation practices built on qualitative data should complement quantitative validation practices (e.g., psychometric analysis). The integration and formal consideration of qualitative methods, alone or in combination with quantitative methods, is an important dimension of validity as a social imperative and reflects current and emerging validation practices in HPE.

What this is doing is it's taking it a step further in saying those forms of evidence can be quantitative and qualitative and I think that's right, absolutely right. (E3)

For users, the current implementation of competency-by-design in medicine program in Canada could facilitate the application of the concept of validity as a social imperative. More specifically, participants (FoG1, FoG4, FoG5) emphasised the coherence of the concept with the changes made in the various training programs such as a programmatic approach or the use of Entrustable Professional Activity (EPAs) in medical education.

² In the text and quotes, participants are identified by the following legend:

E: Interviews with users (educators and committee members)

FoG: Focus group with users (educators and committee members)

L: Interviews with leaders

Required clarification of terms used in the concept analysis

Participants thought some wording used to describe validity as a social imperative could result in confusion or inconsistent interpretations. The term identified as most likely to lead to varied interpretations was the term “society”. During the interviews, all members of both participant groups frequently queried the definition of “society” or “social imperative”, and requested clarification regarding the intended meaning.

For me the concept of society, it is not clear. Because is society my mother, should I prove to my mother that we are finally certifying good physiotherapy students? (FoG3)

Having varied understandings of the intended interpretation of ‘society’ could lead to different uses or application of the concept. For example, in the first characteristic (validity evidence seen as credible by society), we could refer to several different notions of ‘society’; society's needs, evidence to be collected, and different notions of what is considered credible evidence depending on the interpretation of ‘society’.

[...] So society to me is society as a whole. If we're using evidence that is considered credible to society, then validity will become very very challenging to define (L1).

The combination of assessment findings (third characteristic) was understood in two different ways: either as the global interpretation of different assessments instances:

So there's no question that results in a program of assessment, the results of multiple measures and perspectives, do need to be integrated into a whole. And I also agree that integrating doesn't mean just adding up scores on these separate tests (L3).

Or as the global interpretation resulting from more than one person (groups of people making decisions):

So the clinical environment, when you can't make a judgment as an individual practitioner, you go out and you make a collective decision, because it's more complex. (L4).

Participant comments focused on aspects associated with combining assessment data (rather than the combination of validation practices) to obtain a more defensible judgment of a learner's knowledge, attitude, skills or competencies.

Resemblance with modern theories of validity

Participants questioned the notion that the concept was new or emerging. They perceived similarities between the concept of validity as a social imperative and modern validity theories. These similarities may create issues for the acceptability of it as a ‘new’ concept. *Validity built into the assessment process* was seen as similar to modern theories of validity (L1, L2, FoG2).

But both of those [Messick and Kane’s theories of validity] take into account consequential validity and sort of impact on society. So I think at a minimum it’s important to keep in mind that the three discourses [referring to those described by St-Onge et al. [8]] have blurry boundaries (E2).

Leaders agreed that validation is an ongoing process throughout the development and validation stages. For two participants (E3 and E6), validation embedded through the assessment process was to be expected, and one participant (E4) felt that validation built into the assessment process (second characteristic) was coherent with purposefully programmatic assessment, and contributes to credibility and defensibility of the validation process.

In addition to similarities, participants identified strengths which distinguish the concept of validity as a social imperative from other conceptualisations of validity. For example, they found the social role of the validity more explicit (FoG2, FoG5, E5) and they highlighted the importance given to anticipation of consequences (FoG5) in the concept of validity as a social imperative.

Challenges in the application of the concept for practice

Participants anticipated several potential barriers and facilitators associated with the feasibility of the concept of validity as a social imperative, with particular focus on the operationalisation, implementation, and uptake of the concept.

Some participants were uncertain that the concept of validity as a social imperative would facilitate or compound the work expected of comprehensive validation practices.

So if there isn’t a case made, the danger is that people might look at this and say this is just adding a whole other layer of complexity and I don’t see the benefit. (L5)

A leader (L6) expressed doubts about the willingness of medical educators and assessment administrators to put in the time and effort to complete a validation process that aligned with the concept of validity as a social imperative. Educators listed several challenges concerning the limits of their own context such as implicit or explicit institutional values. More specifically, the anticipated costs, time and effort associated with the implementation of the conceptualisation of validity as a social imperative, although currently unknown, was assumed to be a potential limit to its applicability.

It's a bit of a creation of work. They [assessors] kind of want to buy something and just use it and so there is an implication here of time, cost, money, people's time. So if you got too carried away with this, that could be an extra burden. A school's education programs don't have a lot of time or money (L3).

It was difficult for participants to explicitly name what they felt would be needed to apply the concept.

Because we know that we may not create the best assessment for validity and then the responsibility towards the population of what kind of professionals we train. We are conscious of this, but at some point, we have to survive as faculty members. Creating exams is not our only task (FoG5)

Differences between the two stakeholder groups

There were two areas in which the different stakeholder groups appeared to be using lenses that foreground different dimensions of the same underlying themes. Specifically, both groups discussed 1) differing conceptualisations of the importance of assessment and validity and 2) society as a driving force for different ends.

Differing conceptualisations of the importance of assessment and validity

Educators expressed that, through taking part in the study, they were given the opportunity to reflect on the social responsibility of a professor. Interviewees described their awareness of the impact of assessment on the learner and society:

I've never heard of that, but now that I've heard about it, I wonder, 'God, why did it take so long before thinking about this, it's so obvious that we need to do things [assessment] that are valid, it has consequences for society' (FoG4)

Leaders reported appreciating that within the concept of validity as a social imperative there is a consideration for assessors' and administrators' social responsibility regarding all components of an assessment and validation process.

So it [validity as a social imperative] sits well with me because it's saying you have a social responsibility to make a compelling, a judicious argument, about validity. The final product but also there's a process of assessment (L5).

Society as a driving force for different ends

Educators and committee members expressed the importance of the quality of the assessment in order to accurately judge the performance of the future health professionals and thus ensure public safety. They clearly state the value that society places on university degree that serves as proof of the competence of the professionals, and therefore their role in ensuring that a university degree does reflect competence.

Some leaders considered that first characteristic of the concept (validity evidence seen as credible by society) to be the 'added value' of this emerging concept. They discussed that the role of a program or a high-stake exam is to certify the competence of the future health professional – a decision that has a large impact on society.

So I think it's the relevance in terms of (...) some method that can stand up and tell society, look: We believe in these individuals that they are competent, that we are there, like we have that belief in them (L4).

DISCUSSION

The apparent evolution of validity as a social imperative stems from a pragmatic need - the difficulties of using existing validation practices in the current context of assessment in HPE [21–25]. Participants recognised these difficulties and acknowledged the need to bridge the gap between current assessment practices and existing validity theories, and that validity as a social imperative may be an acceptable approach. However, participants asked for clarification of some aspects in the concept of validity as a social imperative (acceptability and feasibility), and anticipated challenges in applying the concept (feasibility).

The perceived lack of clarity appeared concerning for participants; two key concepts seem to generate hesitation regarding the acceptability (uptake by the community) and feasibility (anticipated applicability) of the concept of validity as a social imperative. First, there was a need to better understand the similarities and differences between validity as a social imperative and other conceptualisations of validity. Messick initially described the need to document consequential evidences during a validation process in the *Unified theory of validity* [26,27]. Kane [28,29] - in the *Argument-based validation framework* - argued that consequences of assessment are important to the validation process to support inferences. Recently, Cook and Lineberry [30] suggested that consequential validity is the most important evidence to consider in a validation process. How validity as a social imperative is different - or adds to - existing conceptualisations of validity is the necessity of anticipating consequences throughout the whole assessment process and not only measure consequences *a posteriori*, extending traditional descriptions of consequences. In validity as a social imperative we re-emphasise the importance of consequences of assessment, which are too often neglected in validation processes as supported by Cook et al. [31] and Labbé et al. [32] who found that few studies measured response processes or consequences of the assessment scores in the context of assessment in different HPE contexts.

Second, there is a need to further define "social" and "society" in the concept of validity as a social imperative. A consideration for society in assessment and validity comes with several challenges – the first of which being the way in which we define ‘society’. Cook and Lineberry [30] consider all those affected by the consequences of validity, including “*learners, educators, and educational institutions; patients, providers, and health care institutions; and even society at large*” [30]^(p788). For our participants, the notion of ‘society’ in validity as a social imperative was under-specified. This under-specification may permit for local contextual factors to drive the identification of the relevant consideration of ‘society’. While this may contribute to a non-uniform operationalisation of validity as a social imperative across contexts, this isn’t necessarily a bad thing – varied interpretations of society could be beneficial. To support transparency, we recommend that the relevant “society” be articulated and justified when conducting a validation study.

In order to better understand common uses of the term ‘social’, we examined literature describing social accountability to explore how ‘social’ is typically defined or understood. The World Health Organisation (WHO) [33] has specified that medical schools are accountable to health consumers, health authorities and graduates. Hanson, Rosenberg and Lane [34] describe social responsibility as directed towards patients, families and society as a whole. The Royal College of Physicians and Surgeons of Canada [35] refers to the accountability of the physicians to patients, communities, and broader populations they serve. Communities' and society's needs evolve over time [36], which necessitates evolutions in health care delivery, and consequently in health professional education. The needs of ‘society’ as reflected in social accountability are constantly changing. The word society is flexibly defined in discussions of social accountability, it is not surprising that there was a varied understanding of “social” for our participants when discussing validity as a social imperative.

Both groups of stakeholders anticipated challenges concerning the application of the concept into practice (e.g. time, effort, and cost). This uncertainty regarding resource commitment may be because the concept of validity as a social imperative has not yet been formally operationalised for use in the way that the Standards [32] operationalised Messick’s [27] validity theory, and remains an important next step. In addition, the challenges highlighted by our participants are consistent with Sidani and Braden's framework [18] concerning the feasibility of an intervention (i.e. application of the concept in our study). Indeed, according to these authors [18], the factors influencing the feasibility are, for instance, the context (including the physical and social environment), the resources available, and the training of the individuals involved. Our findings also mirror those of Onyura et al. [38] who identified factors hindering knowledge translation in medical education, with common barriers being the fear of work, role overload, and financial and human resource limitations.

Strengths and limitations

Through purposeful and snowball sampling, we interviewed participants from various backgrounds, organisations, and programs, which contribute to diversity of points of view reflected in our data. Triangulation from different perspectives, an inductive approach and the open-ended questions provided us with rich and varied data. In addition, we relied on Sidani and Braden's [18] framework (e.g. including questions of the relevance of the concept or the applicability to practice). Interpretive description approach [10] provided a rigorous method for analysing, synthesising and transforming data. Co-coding by two team members ensured both the credibility and confirmability of the data [10,14,39]. Thick and rich description of the research process and the sample provided in the manuscript enables readers to establish the transferability to their context [10]. Verification of conclusions with leaders and research team improved credibility and confirmability [20,39].

This study has limitations. We recognise the possible presence of selection bias—participants being employed in Canadian institutions may have limited our sample. However, we interviewed individuals from different disciplines, universities and background. Most of the participants work in medicine and nursing, and it may have been beneficial to have deliberately approached participants involved in other HPE programs. The research team are the same individuals who described the concept of validity as a social imperative in a previous concept analysis, and this may induce some researcher bias. For the leaders, we attempted to minimise this by relying on interviewers who did not know the participants, nor were involved in the original research work describing validity as a social imperative. Further, some participants were critical about validity as a social imperative, suggesting social desirability bias did not appear to influence our data.

Implications

As health professionals' users and leaders, this concept expands our understanding of validation practices that consider the impacts of assessment and validity on society and on future health professionals. Furthering our understanding of how to engage with validity practices adapted to our current assessment approaches in order to improve the quality of

assessment is relevant for several actors including learners, educators, and patients. We believe that the findings from this study will allow us operationalise the concept for use. Identifying potential or perceived barriers to implementation before finalising the operationalisation of the concept of validity as a social imperative enables us to mitigate the perceived difficulties and ensure the concept is accessible to users and leaders.

CONCLUSION

The concept of validity as a social imperative appears to resonate with stakeholders' group (users and leaders in HPE) and reinforces discussions in the literature regarding moving beyond the traditional focus on psychometrics for validation practices reported in HPE. Participants identified several avenues for further refinement of the concept and these remain important areas for future research, including the operationalisation of validity as a social imperative for use in assessment validation in HPE.

Acknowledgments: The authors would like to thank Kathleen Day and Lesley Ananny, Faculty of Medicine, University of Ottawa, Ontario, Canada, who conducted the interviews, and the participants for their generosity and availability.

Conflicts of interest: None.

Appendix 1

Interview guide used for users in a qualitative study to explore perceived acceptability and anticipated feasibility of the concept of validity as a social imperative, 2016

1.	What is your name and what is your role at the University of <...>
2.	<p>Would you tell me about your vision of the validation process (process that measure quality of assessment) in your course or your program?</p> <p>Probes: Can you explain your point of view? What brings you to ...? Can you give me an example?</p>
3.	<p>What do you think of the characteristics that we identified to describe validity as social imperative? (Discuss one by one)</p> <p>Probes: Can you explain your point of view? What brings you to ...? Can you give me an example? What do you like about this characteristic?" What don't you like about this characteristic?</p>
4.	<p>Overall, what is your personal opinion of validity as social imperative?</p> <p>Probes: Can you explain your point of view? What brings you to ...? Can you give me an example? What do you like about the concept?" What don't you like about the concept?</p>

Appendix 2

Interview guide used for leaders in a qualitative study to explore perceived acceptability and anticipated feasibility of the concept of validity as a social imperative, 2017

1.	Can you tell me what motivated you to participate in this study?
2.	<p>Would you tell me what validity in the context of assessment means to you?</p> <p>Probes: Can you explain your point of view? What brings you to ...? Can you give me an example?</p>
3.	<p>What do you think of the characteristics that we identified to describe validity as social imperative? (Discuss one by one)</p> <p>Probes: Can you explain your point of view? What brings you to ...? Can you give me an example? What do you like about this characteristic?" What don't you like about this characteristic?</p>
4.	<p>Overall, what is your personal opinion of validity as social imperative?</p> <p>Probes: Can you explain your point of view? What brings you to ...? Can you give me an example? What do you like about the concept?" What don't you like about the concept?</p>
5.	<p>Do you feel that we have missed important aspects of validity as a social imperative?</p> <p>Probes: Do you have anything else you would like to add? Something else we should consider as we move this work forward? What else can you say about that? Can you explain your point of view? Can you build on that?</p>

References

1. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. *Acad Med*. 2016;91(10):1359–69.
2. Govaerts M, van der Vleuten CPM. Validity in work-based assessment: Expanding our horizons. *Med Educ*. 2013;47(12):1164–74.
3. Govaerts MJB, Schuwirth LWT, van der Vleuten CPM, Muijtjens AM. Workplace-based assessment: Effects of rater expertise. *Adv Heal Sci Educ*. 2011 May;16(2):151–65.
4. Van der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205–14.
5. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478–85.
6. Brown KK, Maryman J, Collins T. An Evaluation of a Competency-Based Public Health Training Program for Public Health Professionals in Kansas. *J Public Heal Manag Pract*. 2017;23(5):447–53.
7. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. *Med Educ*. 2018;52(6):641–53.
8. St-Onge C, Young M, Eva KW, Hodges B. Validity: One word with a plurality of meanings. *Adv Heal Sci Educ*. 2017;22(4):853–67.
9. Thorne S, Reimer Kirkham S, O’Flynn-Magee K. The Analytic Challenge in Interpretive Description. *Int J Qual Methods*. 2004;3(1):1–11.
10. Thorne S. *Interpretative description: qualitative research for applied practice*. New York. Routledge. 2016.
11. Hunt MR. Strengths and challenges in the use of interpretive description: reflections arising from a study of the moral experience of health professionals in humanitarian work. *Qual Health Res*. 2009 Sep;19(9):1284–92.
12. Young M, St-Onge C, Xiao J, Vachon Lachiver E, Torabi N. Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspect Med Educ*. 2018;7(3):182–91.
13. Doja A, Horsley T, Sampson M. Productivity in medical education research: An examination of countries of origin. *BMC Med Educ*. 2014;14(1):1–9.
14. Patton MQ. *Qualitative Research & Evaluation Methods*. 4th ed. Thousand Oaks: Sage Publications; 2015.
15. Brinkman S, Kvale S. *InterViews : learning the craft of qualitative research interviewing*. 3rd ed. Los Angeles. Sage Publications; 2015.
16. Krueger RA, Casey MA. *Focus groups: A practical Guide for Applied Research*. 5th ed. Thousand Oaks: Sage Publications; 2015.
17. Miles MB, Huberman AM, Saldaña J. *Qualitative data analysis*. Sage; 2014.
18. Sidani S, Braden CJCJJ. *Design, Evaluation, and Translation of Nursing Interventions*. Wiley-Blackwell; 2011.
19. Birt L, Scott S, Cavers D, Campbell C, Walter F. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qual Health Res*.

- 2016;26(13):1802–11.
20. Varpio L, Ajjawi R, Monrouxe LV, O’Brien BC, Rees CE. Shedding the cobra effect: Problematising thematic emergence, triangulation, saturation and member checking. *Med Educ*. 2017;51(1):40–50.
 21. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: Assessors’ perspectives. *Adv Heal Sci Educ*. 2013;18(4):559–71.
 22. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the “black box” differently: Assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055–68.
 23. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Acad Med*. 2010;85(5):780–6.
 24. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564–8.
 25. Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ*. 2006;40(4):296–300.
 26. Messick S. Validity. In: Linn RL, editor. *Educational Measurement*. 3rd ed. New York: Macmillan; 1989. p. 13–103.
 27. Messick S. Standards of Validity and the Validity of Standards in Performance Assessment. *Educ Meas Issues Pract*. 1995;14(4):5–8.
 28. Kane MT. Validation. In: Brennan RL, editor. *Educational Measurement*. 4th ed. Westport, CT: American Council on Education/Praeger; 2006. p. 17–64.
 29. Kane M. The Argument-Based Approach to Validation. *School Psych Rev*. 2013;42(4).
 30. Cook DA, Lineberry M. Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Acad Med*. 2016;91(6):785–95.
 31. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Heal Sci Educ*. 2014;19(2):233–50.
 32. Labbé M, Young M, Nguyen LHPP. Validity evidence as a key marker of quality of technical skill assessment in OTL-HNS. *Laryngoscope*. 2018; 128:2296-2300.
 33. Boelen C, Heck JE, World Health Organization. *Defining and measuring the social accountability of medical schools*. Geneva; 1995.
 34. Hanson JL, Rosenberg A a., Lane JL. Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Front Psychol*. 2013;4:668.
 35. Royal College of Physicians and Surgeons of Canada. *CanMEDS 2015 Physician Competency Framework*. Frank JR, Snell L, Sherbino J, editors. Ottawa, Ontario: Royal College of Physicians and Surgeons of Canada; 2015.
 36. Royal College of Physicians and Surgeons of Canada. *Competence by Design (CBD)*.2014; Retrieved from http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/cbd/what_is_cbd_e.pdf
 37. AERA, APA, & NCME. (1999). Validity. In *Standards for Educational and Psychological Testing* (pp. 9–24). Washington: American Educational Research Association.

38. Onyura B, Légaré F, Baker L, Reeves S, Rosenfield J, Kitto S, et al. Affordances of Knowledge Translation in Medical Education. *Acad Med.* 2015;90(4):518–24.
39. Morse JM. Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qual Health Res.* 2015;25(9):1212–22.

5.2 Compléments méthodologiques

Nous avons interrogé deux profils de participants soit 1) le personnel enseignant et les membres de comités d'évaluation des apprentissages et 2) les personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada. Dans la section suivante, nous précisons les raisons des choix méthodologiques pour chacun des différents profils.

5.2.1 Personnel enseignant et membres de comités d'évaluation des apprentissages

Les participants du profil « personnel enseignant et membres de comités d'évaluation des apprentissages » œuvrent dans un programme universitaire professionnalisant en santé au Québec. Plus précisément, elles sont impliquées dans les programmes de médecine, de sciences infirmières, d'ergothérapie ou de physiothérapie et travaillent dans une des universités québécoises qui offrent ces programmes. Il s'agit de l'Université McGill, l'Université de Sherbrooke, l'Université de Montréal et l'Université Laval. Les membres du personnel enseignant et de comités d'évaluation sont essentiels à cette recherche puisqu'ils possèdent une expérience concrète en ce qui a trait à la conception, la validation et le monitoring de méthodes d'évaluation des apprentissages. Ils représentent les utilisateurs potentiels du concept de validité en tant qu'impératif social.

Afin d'être inclus dans l'étude, les participants devaient être des professeurs, des chargés de cours ou des chargés d'enseignement ou encore des membres de comités d'évaluation des apprentissages possédant une expérience de plus d'un an. Les personnes ciblées devaient travailler dans un des programmes professionnalisants en santé ciblés (médecine, sciences infirmières, ergothérapie ou physiothérapie) et œuvrer dans une université offrant les quatre programmes identifiés précédemment. Les personnes qui n'enseignaient que quelques heures dans un cours, les tuteurs, les moniteurs ou les superviseurs de stage ont été exclus de l'étude.

Recrutement

Nous avons recruté le personnel enseignant et les membres de comités d'évaluation des apprentissages en utilisant une méthode d'échantillonnage par choix raisonné. Afin

d'explorer différentes perspectives, nous avons ciblé des personnes provenant de chacune des quatre universités et œuvrant dans les quatre programmes de formation identifiés. Dans un premier temps, l'étudiante a communiqué par courriel avec chacun des directeurs des programmes de médecine, de sciences infirmières, d'ergothérapie et de physiothérapie dans les quatre universités ciblées par la recherche. Le contenu du courriel (annexe 5) expliquait l'objectif de la recherche, le profil des participants recherchés ainsi que la description de la nature de la participation. Dans un deuxième temps, le directeur de programme partageait notre invitation aux personnes répondant aux critères d'inclusion. Dans un troisième temps, les personnes intéressées par la recherche communiquaient avec nous afin d'obtenir des explications supplémentaires sur l'étude. Après ce premier contact, nous déterminions la plage horaire et le lieu des entrevues selon leurs disponibilités.

Collecte des données

Lors de la planification de la recherche, nous avons envisagé mener quatre groupes de discussion focalisée, soit un groupe par université. Un avantage des entrevues de groupe résidait dans la richesse des échanges et l'interaction entre les participants (Beck, 2013; Krueger & Casey, 2015; Morgan, 2019; Thorne, 2016) au sujet de la validité en tant qu'impératif social. Nous avons choisi de regrouper de cinq à huit personnes (Beck, 2013; Krueger & Casey, 2015; Morgan, 2019) au sein d'une même université, sans égard pour leur discipline professionnelle. Toutefois, pour rejoindre le plus de personnes possible, nous avons également mené des entrevues individuelles auprès de quelques personnes non disponibles au moment de la rencontre de groupe dans leur milieu d'enseignement. Idéalement, lorsque c'était possible, nous avons privilégié les entrevues de groupe.

Un questionnaire sociodémographique (annexe 6) a été rempli par écrit au moment de l'entrevue de groupe ou individuelle. Le questionnaire était composé de neuf questions. Six items portaient sur des éléments plus généraux (genre, groupe d'âge, discipline professionnelle et emploi à l'université) afin de décrire l'échantillon. Trois questions avaient pour objectif de préciser la contribution de la personne participante dans le domaine de l'évaluation des apprentissages.

L'étudiante a construit un guide d'entrevue semi-structuré composé de quatre questions pour les entrevues de groupes de discussion focalisée (annexe 7) et pour les entrevues individuelles (annexes 8). Le guide a été ensuite validé par l'équipe de recherche. Nous avons privilégié l'utilisation de questions ouvertes et de questions de clarification (*probes*) pour explorer le point de vue des participants (Tracy, 2013). Les questions portaient principalement sur l'acceptabilité et la faisabilité potentielles de l'application du concept de validité en tant qu'impératif social. Inspirées des travaux de Sidani et Braden (2011), les questions du guide d'entrevue semi-structuré ont permis de mieux comprendre la perception des personnes interviewées au sujet de l'acceptabilité et de la faisabilité anticipée de l'application du concept de validité en tant qu'impératif social. Les questions sur l'acceptabilité portaient sur les thèmes suivants : 1) la pertinence; 2) la commodité; 3) l'efficacité, 4) les risques associés et 5) l'adhésion à cette vision de la validité de l'évaluation des apprentissages (Sidani & Braden, 2011). Les questions sur la faisabilité abordaient les thèmes suivants : 1) la disponibilité et la qualité des personnes évaluatrices; 2) la formation des personnes évaluatrices; 3) les ressources nécessaires; 4) le contexte; 5) la fidélité et 6) la portée (Sidani & Braden, 2011). Le caractère semi-structuré du guide d'entrevue favorisait l'ajout de questions au besoin.

Déroulement

Le déroulement était similaire pour les entrevues de groupes et les entrevues individuelles. Les personnes ayant manifesté un intérêt ont reçu par courriel la confirmation de l'heure et du lieu de l'entrevue. Les entrevues de groupe et individuelles se sont déroulées en octobre et novembre 2016 dans les milieux d'enseignement des personnes recrutées. Lorsque les entrevues se déroulaient à l'extérieur du milieu connu de l'étudiante, les participants étaient invités à communiquer les informations nécessaires à la réservation d'un local pour la tenue des entrevues.

Deux documents ont été remis par courriel aux participants avant l'entrevue de groupe ou l'entrevue individuelle. Les documents étaient disponibles en français ou en anglais, selon la préférence de la personne. Un premier document (annexe 9) de deux pages résumait les résultats de l'analyse de discours effectuée par St-Onge et al. (2017). Ces auteurs ont décrit

trois conceptualisations de la validité dans le domaine de l'évaluation des apprentissages et parmi celles-ci, on retrouve une première description de la validité en tant qu'impératif social. Un deuxième document (annexe 10) de deux pages présentait une synthèse des résultats de l'analyse de concept (phase 1 de la recherche). Les participants devaient examiner ces documents avant l'entrevue afin de favoriser l'expression de leurs idées, leur point de vue au sujet de l'acceptabilité et la faisabilité anticipée des résultats de l'analyse de concept.

Pendant les entrevues de groupes de discussion focalisée, une animatrice (l'étudiante) a interviewé les participants alors qu'une observatrice (une directrice de recherche) notait dans un journal de bord des commentaires à propos des thèmes abordés, de l'atmosphère et des interactions entre les participants (Krueger & Casey, 2015). Au besoin, l'observatrice ajoutait des questions pour explorer plus en profondeur des idées exprimées. Pour les entrevues individuelles, l'intervieweuse était seule avec la personne participante, comme c'est habituellement le cas pour les entrevues individuelles en recherche qualitative. Les entrevues ont été menées en anglais ou en français selon la préférence de la personne ou du groupe interviewés. Avant le début de l'entrevue, l'animatrice s'assurait que le questionnaire sociodémographique et le formulaire de consentement (annexes 11 et 12) étaient remplis par les personnes présentes. D'une durée variant de 60 à 90 minutes, toutes les entrevues ont été enregistrées pour faciliter la transcription et l'analyse des données (Brinkman & Kvale, 2015).

5.2.2 Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada

Nous avons interrogé des personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada. Ces personnes influencent considérablement les changements de pratique dans le domaine de la pédagogie des sciences de la santé, que ce soit auprès des organismes accréditeurs ou auprès des établissements d'enseignement. Pour être incluses, les personnes devaient exercer ou avoir exercé une influence dans le domaine de la validité ou de l'évaluation des apprentissages dans le domaine des sciences de la santé. Par exemple, les personnes pouvaient être sélectionnées par l'équipe de recherche en raison de

leur implication passée ou actuelle dans les comités portant sur l'évaluation des apprenants auprès des organismes accréditeurs. Les personnes pouvaient également être sollicitées si elles avaient publié plusieurs articles pertinents sur le sujet, articles qui avaient été révisés par des pairs. Nous n'avons pas déterminé de critères d'exclusion.

Recrutement

Nous avons utilisé une méthode d'échantillonnage par choix raisonné et par boule de neige pour recruter les participants. Nous avons délibérément sélectionné les premières personnes participantes afin d'obtenir une variété de profils (p. ex. discipline professionnelle, lieu géographique, expérience diversifiée). Les personnes recrutées par l'étudiante ont ensuite recommandé des collègues répondant aux critères d'inclusion. Un premier courriel a été envoyé aux personnes ciblées pour expliquer la recherche et les convier à participer à l'étude. Les personnes intéressées communiquaient ensuite avec une assistante de recherche pour déterminer le meilleur moment pour l'entrevue.

Collecte des données

Afin de nous adapter aux disponibilités et au lieu de résidence des personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada, nous avons décidé de mener des entrevues individuelles par téléphone. Nous avions prévu obtenir la participation de cinq à huit personnes. L'entrevue individuelle comporte plusieurs avantages tels qu'une plus grande flexibilité concernant les disponibilités ou encore plus de précision dans les réponses de la personne participante (Patton, 2015; Thorne, 2016). L'animatrice de l'entrevue a la possibilité de clarifier les propos de la personne interrogée.

Avant l'entrevue, les participants recevaient un courriel (annexe 13) avec un hyperlien, permettant de remplir en ligne un questionnaire sociodémographique (annexe 14) et un formulaire de consentement (annexe 15). Le questionnaire comprenait six questions portant sur le genre, le groupe d'âge, leur formation initiale et leur expérience dans le domaine de l'évaluation des apprentissages et de la validité. Composé de cinq questions ouvertes, le guide d'entrevue semi-structuré a été construit par l'étudiante et adapté au profil de participant (annexe 16). Les thèmes explorés étaient similaires à ceux abordés lors

des entrevues avec le personnel enseignant et les membres de comités d'évaluation des apprentissages, soit l'acceptabilité et la faisabilité anticipée de l'application du concept de validité en tant qu'impératif social.

Déroulement

Les participants ont reçu par courriel deux documents préalables à l'entrevue individuelle (annexes 17 et 18). Il s'agissait des mêmes documents que pour le personnel enseignant et les membres de comités d'évaluation des apprentissages. Les documents étaient remis en anglais, langue utilisée par les participants.

Une professionnelle de recherche a été sélectionnée par l'équipe de recherche pour mener les entrevues individuelles en anglais avec les personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada. Celle-ci possédait une expérience pertinente en pédagogie des sciences de la santé et dans l'animation d'entrevues. Comme la majorité des participants connaissaient les directrices du projet doctoral, nous croyons que l'implication d'une personne extérieure au projet (professionnelle de recherche) a diminué le biais de désirabilité sociale et a facilité l'expression de critiques envers le concept de validité en tant qu'impératif social. Toutefois, comme la professionnelle de recherche n'avait pas une connaissance approfondie du sujet de la recherche, elle a reçu de la rétroaction de l'étudiante et de l'équipe de recherche après la première et la deuxième entrevue. Cette rétroaction a assuré la qualité des échanges entre l'animatrice et les participants. D'une durée de 60 à 90 minutes, les entrevues téléphoniques se sont déroulées entre janvier 2017 et juillet 2017. Les entrevues ont été enregistrées et transcrites de manière anonyme.

5.3 Caractéristiques sociodémographiques complémentaires

Les résultats provenant des entrevues et des groupes de discussion focalisée ont été explicités dans l'article 2. Dans cette section, nous compléterons la présentation des caractéristiques sociodémographiques pour les deux profils de participants.

5.3.1 *Personnel enseignant et membres de comités d'évaluation des apprentissages*

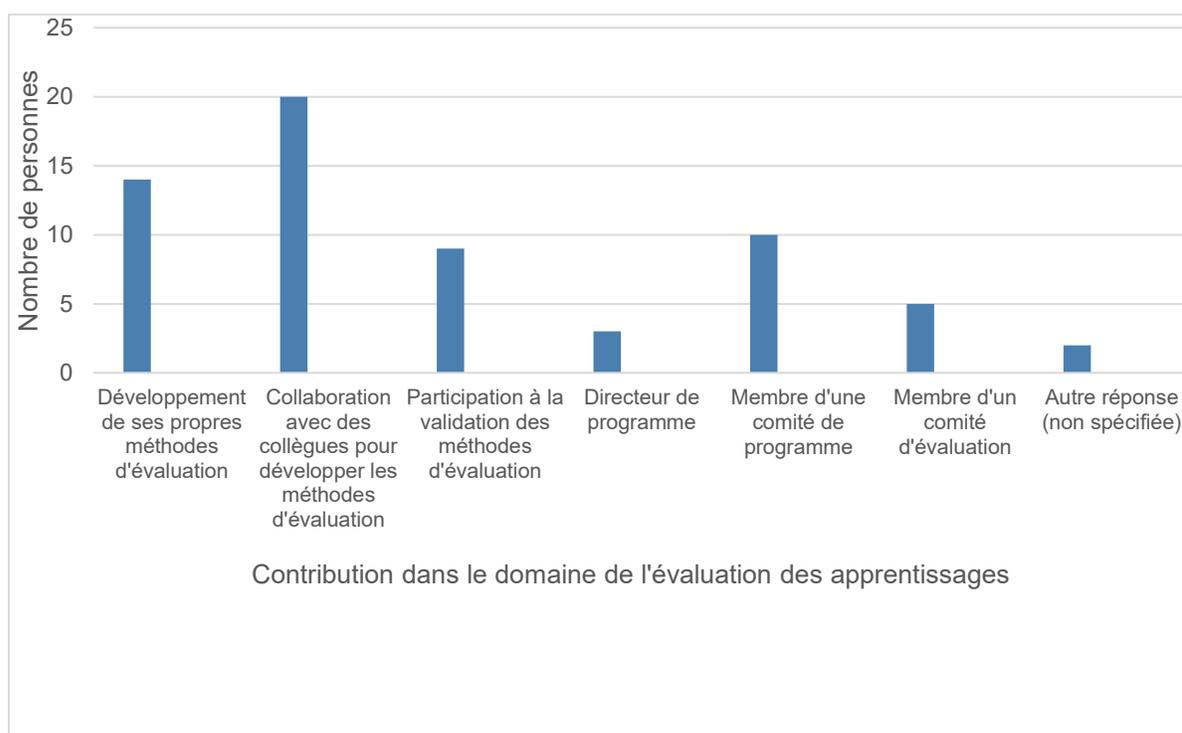
Nous avons rencontré un total de 23 personnes provenant de quatre universités du Québec (Université de Sherbrooke, Université de Montréal, Université McGill et Université Laval). Plus précisément, nous avons mené cinq groupes de discussion focalisée (n=19 personnes) et quatre entrevues individuelles. Les caractéristiques sociodémographiques sont décrites dans le tableau 10. L'échantillon était composé à parts égales d'hommes et de femmes provenant de différentes disciplines : physiothérapie, ergothérapie, médecine, sciences infirmières et éducation. La majorité des personnes interrogées occupaient alors un emploi de professeur ou de professeur de clinique. Les autres personnes travaillaient comme chargés de cours ou chargés d'enseignement clinique ou comme conseillers pédagogiques. Quatre personnes nous ont mentionné exercer plus d'une fonction simultanément (par exemple, chargé de cours et conseiller pédagogique). Dans le cadre de leur emploi, les participants s'impliquaient au premier cycle universitaire, au deuxième cycle universitaire ou encore avaient une tâche à la fois au premier et au deuxième cycle.

Tableau 10
Caractéristiques sociodémographiques
du personnel enseignant et membres de comités d'évaluation des apprentissages
N = 23

Caractéristiques	Nombre (%)
<i>Sexe</i>	
Homme	12 (52 %)
Femme	11 (48 %)
<i>Discipline</i>	
Sciences infirmières	8 (35 %)
Médecine	7 (30 %)
Physiothérapie	2 (9 %)
Ergothérapie	2 (9 %)
Éducation	4 (17 %)
<i>Emploi</i>	
Professeur ou professeur de clinique	16 (70 %)
Chargé d'enseignement ou chargé de cours	5 (22 %)
Conseiller pédagogique	2 (9 %)
<i>Implication universitaire</i>	
1 ^{er} cycle	4 (17 %)
2 ^e cycle	5 (22 %)
1 ^{er} et 2 ^e cycle	14 (61 %)

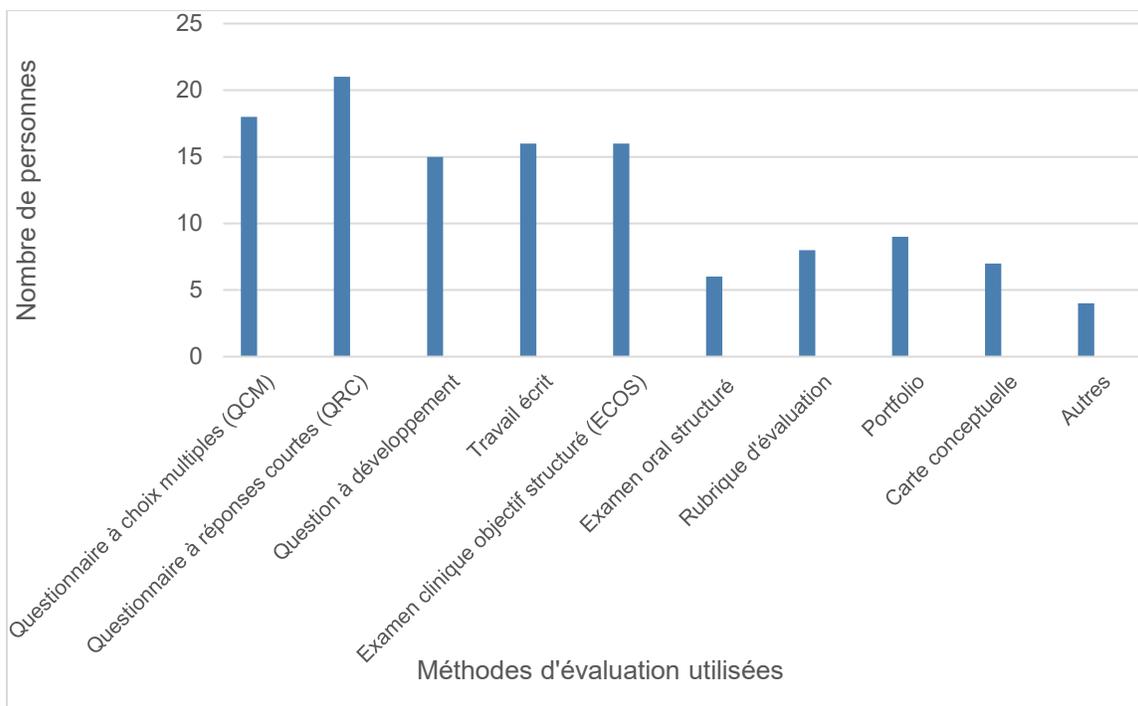
La contribution de ces participants à l'évaluation des apprentissages est illustrée à la figure 3. Nous constatons que la majorité des personnes rencontrées développent leurs propres méthodes d'évaluation et collaborent avec des collègues au développement de méthodes d'évaluation.

Figure 3
Contribution des participants
N = 23



Nous avons questionné les participants sur les stratégies utilisées lors de l'évaluation des apprentissages dans leur cours (ou programme). Les résultats sont présentés dans la figure 4. Les méthodes d'évaluation les plus utilisées sont : les QCM, les questionnaires à réponses courtes (QRC), les questions à développement, le travail écrit et les ECOS.

Figure 4
Méthodes d'évaluation des apprentissages utilisées par les participants
N=23



5.3.2 Personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada

Nous avons rencontré sept personnes ayant accepté de participer à la recherche. Nous avons invité un total de treize personnes. Les personnes ayant refusé de participer ont invoqué le manque de temps. Les caractéristiques sociodémographiques sont décrites dans le tableau 11. La majorité des participants étaient des hommes ($n = 5$) âgés de 45 à 54 ans ($n = 6$) ayant une formation initiale en médecine ou en psychologie. Au moment de la recherche, elles s'impliquaient dans le domaine de la pédagogie médicale ($n = 5$) ou de la pédagogie des sciences de la santé ($n = 2$). Les participants avaient en moyenne 18 ans d'expérience dans leurs rôles actuels.

Tableau 11
Caractéristiques sociodémographiques
des personnes engagées dans la recherche au Canada
N=7

Caractéristiques	Nombre
<i>Sexe</i>	
Homme	5
Femme	2
<i>Âge</i>	
35-44 ans	1
45-54 ans	6
<i>Formation initiale</i>	
Médecine	4
Psychologie	2
Autre	1
<i>Implication et emploi actuels</i>	
Pédagogie médicale	5
Pédagogie des sciences de la santé	2

Comme il s'agissait de personnes engagées dans la recherche en pédagogie des sciences de la santé au Canada, nous souhaitons mieux comprendre leur implication dans le domaine de l'évaluation des apprentissages et de la validité. Le tableau 12 présente les réponses des participants. La majorité des personnes interrogées mènent des travaux sur l'évaluation des apprentissages et sur la validité (n = 6). Elles ont répondu avoir une bonne compréhension des problématiques actuelles dans le domaine de l'évaluation et de la validité.

Tableau 12
Description de l'expérience du participant
N=7

	Totalemment en désaccord	En désaccord	Plutôt en désaccord	Plutôt d'accord	En accord	Totalemment en accord
Je travaille régulièrement sur l'évaluation des apprentissages		1				6
Je travaille régulièrement sur la validité	1				2	4
Dans mon travail, la plupart de mes tâches sont liées aux enjeux de l'évaluation des apprentissages	1		1	2		3
Je pense avoir une bonne compréhension des problématiques actuelles de la validité		1			1	5

6. MÉTHODE, RÉSULTATS ET DISCUSSION DE LA PHASE 3 – ÉLÉMENTS À CONSIDÉRER

Le troisième objectif de la recherche visait à décrire les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social. Dans ce chapitre, nous présenterons la méthode, les résultats et la discussion sous la forme d'un article et de données complémentaires.

6.1 Article 3

How to put social into validity: Guideposts for considerations

Auteurs de l'article

Marceau, Mélanie; Young, Meredith; Gallagher, Frances; St-Onge, Christina

Statut de l'article

Soumis le 9 août 2019 à *Academic Medicine*

Avant-propos de l'article

Cet article présente les résultats de l'objectif 3 de l'étude doctorale qui vise à décrire les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social. L'article n'est pas écrit sous la forme classique (c.-à-d. introduction, méthode, résultats et discussion) mais plutôt sous la forme d'une perspective. Dans la revue *Academic Medicine*, un article de type « perspective » permet aux auteurs de décrire un point de vue sur une problématique en pédagogie des sciences de la santé et de proposer une hypothèse ou une piste de solution appuyée par les écrits scientifiques.

Contribution de l'étudiante et des coauteurs

Les directrices de recherche ont supervisé l'ensemble des travaux en plus de contribuer à la rédaction de l'article. L'étudiante a regroupé dans un fichier Excel l'ensemble des résultats découlant des objectifs 1 et 2. Par la suite, les données qualitatives ont été coanalysées par l'étudiante et ses directrices. Les tableaux ont été présentés et discutés à plusieurs reprises avec l'ensemble de l'équipe jusqu'à l'obtention d'un consensus. L'étudiante a rédigé le premier jet du manuscrit et a procédé au raffinement du contenu, qui a été bonifié par les réflexions et commentaires de l'ensemble des auteurs.

How to put social into validity: Guideposts for consideration

Mélanie Marceau, MSc, PhD candidate, Meredith Young, PhD, Frances Gallagher, PhD, and Christina St-Onge, PhD.

M. Marceau is a PhD candidate, Department of Nursing, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada.

M. Young is Associate professor, Department of Medicine and Institute for Health Sciences Education, Faculty of Medicine, McGill University, Montreal, Québec, Canada.

F. Gallagher is Professor, Department of Nursing, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada.

C. St-Onge is Professor, Department of Medicine, Faculty of Medicine and Health Sciences, Université de Sherbrooke, Sherbrooke, Québec, Canada. She holds the *Paul Grand'Maison de la Société des médecins* Research Chair in Medical Education.

Correspondence should be addressed to: Mélanie Marceau, Faculty of Medicine and Health Sciences, Université de Sherbrooke. 3001 12th Avenue North, Sherbrooke, Québec, Canada, QC J1H 5H4

Tel: 00 1 819 821 8000; 72435

E-mail: melanie.marceau@usherbrooke.ca

Résumé

Face à l'évolution des pratiques évaluatives intégrées dans une approche par compétences (APC) telles que les activités professionnelles fiables, les jalons de compétences et l'évaluation programmatique, des incertitudes sont apparues quant aux analyses psychométriques comme indicateurs principaux de la qualité de l'évaluation des apprentissages dans un processus de validation. Par conséquent, une attention particulière a été portée à l'examen de l'influence des pratiques évaluatives sur la conceptualisation, l'opérationnalisation et l'application de la validité.

Les chercheurs et les enseignants réclament des processus de validation qui tiennent compte des répercussions de l'évaluation des apprentissages sur les apprenants et sur la société. Cette préoccupation sociale a été décrite comme étant la validité en tant qu'impératif social. Le concept de validité en tant qu'impératif social se caractérise par la validité intégrée dans le développement et le monitoring de l'évaluation des apprentissages afin de garantir la qualité de l'évaluation et de justifier l'interprétation des résultats tout en portant une attention particulière aux conséquences sociales de l'évaluation.

Étant donné la nouveauté du concept de validité en tant qu'impératif social dans la littérature en pédagogie des sciences de la santé, il est difficile de savoir comment mettre en œuvre les pratiques de validation dans une perspective d'impératif social. Les auteurs proposent ici des éléments à considérer lors de l'application de la validité en tant qu'impératif social au cours du développement, de l'administration et du monitoring de l'évaluation des apprentissages, afin de prendre en compte les conséquences sociales de l'évaluation. Les éléments à considérer prennent la forme de balises pour la mise en œuvre du processus de validation en cohérence avec la validité en tant qu'impératif social - une conceptualisation mieux adaptée aux tendances actuelles en matière d'évaluation, telles que l'évaluation programmatique et longitudinale.

Abstract

With current advances in competency-based assessment (CBA) such as entrustable professional activities (EPAs), milestones, and a focus on programmatic assessment, uncertainties have arisen about the reliance on psychometric markers as primary indicators of assessment quality in a validation process. Thus, renewed attention is being paid to how current assessment trends may influence how validity is conceptualized, operationalized, and applied.

Researchers and educators have called for validation approaches that include consideration for the impact of assessment and score interpretations on individual students and society as a whole. This social emphasis, in validity and validation practices, has been described as *validity as a social imperative*. Descriptions of validity as a social imperative forefront the notion of building quality into the assessment development and monitoring processes as a means to ensure assessment quality and defensibility of score interpretation, with careful attention on the social consequences of assessment.

Given the relative novelty of validity as a social imperative in the HPE literature, it is difficult to know how to operationalize and implement validation practices from a social imperative perspective. Here, the authors propose guideposts for consideration when embedding validity as a social imperative throughout assessment development, administration, and monitoring while attending to the social consequences of assessment. These guideposts propose areas of consideration for implementing validation practices aligned with validity as a social imperative – a conceptualization that may be better suited to current assessment trends such as programmatic and longitudinal assessment programs.

Introduction

Programmatic assessment,^{1,2} authentic assessment,³⁻⁵ and the development of milestones and entrustable professional activities (EPAs)^{6,7} have changed the landscape of assessment in health profession education (HPE). Given that assessments play a role in gatekeeping for learners entering training and practice, it is critical to ensure that defensible decisions are made based on high quality assessment data. Here, we present several guideposts for consideration for how to interweave validity as a social imperative throughout assessment development and quality monitoring.

Validity is a key indicator of the appropriateness of score interpretation. Authors of validity theories (e.g. Messick or Kane) describe validation as a process involving the collection of evidences to support a given interpretation of assessment data.^{8,9} Authors have expressed hesitations about the use and applicability of traditional validation practices for an increasing number of novel approaches to assessment aligned with the competency-based education model.^{5,10-14} Traditional validation and quality monitoring practices rely heavily on psychometric approaches, which may not be well suited for some innovations such as programmatic assessment.^{1,2}

Likely reflecting the mismatch between current assessment trends and available validation approaches, some authors have tackled issues of conceptualization of validity – how it is defined and understood.¹⁵ Others highlighted questions related to the operationalization of validity into validation practices – what is, or should be, the ‘gold standards’ in validation processes.^{1,16-18} Other authors have looked into the application of validity conceptualizations and documenting how these are enacted as validation practices in literature.¹⁹⁻²² There are few documented suggestions as to how to adapt validation models to our current assessment landscape, and this gap between current assessment and validation practices appears to have encouraged reflection regarding validity, validation, and assessment practices in HPE.

Validity of assessment as a social imperative

As a result of the apparent tensions between current assessment and validation practices, there is an emerging conceptualization of validity within HPE - the concept of validity as a social imperative.²³ This conceptualization of validity was first identified in a discourse analysis by St-Onge et al.,¹⁵ and refined by Marceau et al.²³ The concept of validity as a social imperative highlights considerations for the social dimension or impact of validity. We recognize that who is, or can be, considered as 'society' differs depending on the assessment context under consideration. Here, for the purposes of this article, we define society as the stakeholders involved in the assessment process- specifically learners and assessors. The concept of validity as a social imperative forefronts the notion of purposefully building validation into the assessment development and monitoring processes at the outset as a means to increase assessment quality and the defensibility of score interpretation, with careful attention paid to the social consequences of assessment. The concept of validity as a social imperative is composed of four attributes: 1) validity evidence seen as credible by society; 2) validity built into the assessment process; 3) interpretation of the combination of assessment findings and 4) validity evidence included quantitative and qualitative data.²³ (defined in Table 1)

In a follow up study, Marceau, St-Onge, Gallagher and Young²⁴ explored the acceptability and perceived feasibility of validity as a social imperative with individuals engaged in assessment and HPE research. While participants were supportive of the concept, Marceau et al.²⁴-reported that participants questioned *how* the idea of validity as a social imperative could be made accessible to educators and administrators responsible for creating, administering and monitoring assessment within HPE programs.

The concept of validity as a social imperative has only recently been formally described and reflects considerations currently at play in the HPE assessment literature.²³ While considerations for consequences of assessment or mobilization of qualitative data are also present in modern validity theories,^{8,9} several characteristics of validity as a social imperative remain unique. Firstly, validity as a social imperative includes the importance of documenting validity evidences as an explicit component of the assessment development

and administration processes. Secondly, and relatedly, documenting validity evidence does not start once the exam has been administered and assessment data is available for analysis given the consideration of validity evidence beyond psychometric markers of quality. Finally, a key characteristic of validity as a social imperative is the overarching concern for the consequences of assessment for individual learners and society at large throughout the assessment development and monitoring processes.

Guideposts for embedding the social into validation practices

In this paper, we offer suggestions for how the social imperative of validity can be woven throughout the assessment development and monitoring processes (summarized in Table 1; expanded in detail below). In order to articulate how the concept of validity as a social imperative can be considered for use, we first describe and define the attributes of the concept of validity as a social imperative.²³ Then, we describe the suggested guideposts including 1) target stages (i.e. when in the assessment development and validation process it should be considered) and 2) strategies for application.

In order to create the guideposts for consideration, we used Thorne's descriptive interpretive approach.²⁵ As we are choosing to use the term, guideposts can be considered to be a collection of target stages and strategies for applying the concept of validity as a social imperative for use (Table 1; in essence guideposts are suggesting a 'when' and several 'how's'). Analysis was guided by Miles, Huberman and Saldana,²⁶ which consists of three simultaneous analysis cycles (data condensation, data display, drawing and verifying conclusions. For data condensation, two team members (MM and CSTO) conducted qualitative data analysis from a concept analysis²³ and interviews conducted with several stakeholders.²⁴ Codes emerged from an inductive process. Data display facilitated the organization of the data in the way to order and construct the guideposts, structured by attribute. Drawing and verifying conclusions required regular discussions with the whole team until a consensus was reached.

An overarching goal of validity as a social imperative is to document validity evidence for an assessment or assessment program that is seen as credible by society, or in other words,

considered trustworthy by society (table 1; attribute 1). As we described earlier, who is considered as 'society' depends deeply on the assessment and validation context. The relevant definition of who constitutes 'society' in a given assessment and validation process must be defined by the individual or individuals responsible for the targeted assessment or assessment program. For example, for a high stakes' assessment such as a certification examination, society would likely be defined as learners, assessors, patients, and health care institutions. We suggest that appropriate considerations for 1- planning the assessment, 2- planning the administration of the assessment, 3- documenting validity evidences according to a structured and flexible plan and 4- planning interpretation of validity evidence and assessment scores, and documenting these processes, evidences, and decisions can be critical to providing evidence of validity that is seen as credible to society.

Table 1
Attributes and guideposts to consider in validity as a social imperative

Description of the concept of validity as a social imperative		Guideposts	
Attribute	Definition	Target stages:	Strategies for application
Validity evidence seen as credible by society	Professional bodies (teaching institutions and professional orders) must be able to document, in a way that is perceived as credible by society, decisions made regarding a students' academic progress and their level of competence for embarking on an independent professional career.	<i>Planning the overall validation process including assessment, administration, and monitoring, with consideration for stakeholder views and expectations.</i>	
Validity built into the assessment process	Validity evidence includes the justification of decisions made during the development and administration of an assessment, and the interpretation of assessment data. This evidence includes consideration of the potential consequences that the interpretation of the assessment scores could have on the individual, the institution, and society.	<i>Planning the assessment</i>	Prepare a repository or documentation framework for collating validation data. Explicitly document the purpose and stakes of the assessment. Situates individual assessments in the larger educational and assessment plan.
		<i>Planning the administration of the assessment</i>	Anticipate potential consequences for an individual learner, institution, and society, and implement mitigation strategies.
Validity evidence includes quantitative and qualitative data	Validity evidence must be collected using rigorous approaches, and attention should be paid to quantitative and qualitative data sources as legitimate validity evidence.	<i>Documenting validity evidences according to a structured and flexible plan</i>	Identify resources available for validation process. Choose validity evidences strategically, ensuring the presence of qualitative and quantitative data, if appropriate. Adjust the validation process according to the documented evidences of validity
Interpretation of the combination of assessment findings	Assessment data generated within an assessment program are often combined to make a final judgment. Greater emphasis should be given to validity evidences gathered for the overall assessment data interpretation – than evidences for individual tools or assessment episodes when the overall assessment data informs decisions about learners.	<i>Planning the interpretation of validity evidences and assessment scores</i>	Examine quality of scores collected given the intended score use. Critically review the quality of evidence documented through the validation process.

Adapted from Marceau et al.(Marceau et al., 2018) and Marceau et al.(Marceau et al., 2019)

Planning the assessment

Documenting and justifying decisions made throughout the assessment planning and development process is one of the key attributes of validity as a social imperative. The documentation of decisions such as the purpose of an assessment, the construct to be assessed, and the place of the assessment within an educational or assessment program allows educators and administrators the opportunity to identify potential intended or unintended consequences of their assessment. This purposeful planning of assessment and documenting key decision points in the assessment development process, is the starting point of any socially responsible assessment development and validation process.

Prepare a repository or documentation framework for collating validation data

Putting in place a comprehensive approach to purposefully and comprehensively document the anticipated approach to validation is one practice that supports a socially responsible approach to validation. Documentation of validation processes should be done explicitly and clearly, with enough supporting documentation for external, or historical, review. More specifically, tracking decisions and rationales in a logbook will help to keep track of the validation process and validity evidences necessary to support the interpretation of assessment data. Furthermore, to be comprehensive, documentation should include not only a deep description of the validation process but also, how, and why a given decision was made in order to support and help contextualize the future interpretation of assessment scores.

Explicitly document the purpose and stakes of the assessment

An assessment should be “fit for purpose”²⁷⁻³⁰ – meaning that a given assessment or program of assessment should be designed in a way that reflects its intended purpose and use. Purpose and stakes of the assessment must be explicit (i.e. no hidden agenda), and known to all individuals involved in the assessment and validation process (learners, assessors, institutions and professional bodies) as they can influence the interpretation of assessment data.^{10,17,31} For example, assessment with the intention to provide feedback has a different purpose than assessment with the intention to determine readiness for practice. Validation processes should be tailored to the stakes of assessment,³²⁻³⁴ just as they are

tailored to the purpose of assessment – meaning, for example, more evidences should be provided when the assessment is used for high-stakes (such as licensure or certification) versus lower-stakes (such as providing feedback to trainees during a clinical rotation).

Situate individual assessments in the larger educational and assessment plan

Consideration regarding the purpose of assessment in terms of student learning (i.e. testing that learning has happened or whether the assessment is meant to shape or support learning) and the anticipated impact on learning (assessment that drives learning) must guide the choice of assessment strategy.^{35,36} Anticipating the purpose of assessment and intended use of assessment scores should shape the assessment strategy or strategies chosen, in addition to their placement within an assessment program. For example, if considering the potential benefits of assessment for shaping student learning is a priority of a given assessment program, assessments should be frequent and constant and provide sufficient feedback.^{34,37}

When considering the implementation of a program of assessment using various assessment strategies, or having multiple assessment opportunities, can be used to purposefully gather multiple snapshots of a learner.^{4,17,27,29,33,38,39} Like a mosaic, having multiple snapshots likely allows for a better overview of learners' performance.³⁶ Each part of the mosaic (each snapshot) should be chosen wisely, and the combination of assessment strategies should be considered in a way that complements their inherent strengths and weaknesses.^{1,17,27,30,32,38} The combination of assessment snapshots facilitates the triangulation of information about a learner, meaning it provides information about the student's performance from different sources and different contexts.^{36,40,41} Then, decisions resulting from the interpretation of scores will be supported by stronger arguments, as they will be based on the coordination of multiple perspectives, from multiple assessment strategies, across time.

Planning the administration of an assessment

Problems can arise during the administration of an assessment or program of assessment as we will illustrate below. These issues could lead to negative consequences for individuals or for society. Appropriate planning may help avoid or reduce potential issues during administration and therefore minimize subsequent consequences.

Anticipate potential consequences and implement mitigation strategies

Considering what could go wrong during the administration (i.e. delivery) of an assessment and the potential consequences for individuals and society, is a key demonstration of the social imperative of those responsible for a given assessment or program of assessment. Here, we consider both administrative-related considerations, and more human-associated considerations that could influence the quality of assessment data generated. For administrative-related considerations, consequences could be seemingly small - such too few or missing physical copies of an assessment. However, these small administrative issues may increase learner's anxiety³⁴ and impact the quality of assessment data.^{34,42} Care must be taken to have the correct number of exam copies or to check the quality of the print.³⁰ As well, the security of the assessment and assessment scores is also a non-negligible aspect which must also be planned before the test is administered⁴²; as cheating or having unauthorized access to exam content means that scores are no longer a representation of a given students' ability. Once anticipated consequences are identified, they should be minimized to ensure the legitimacy of score interpretation and judgment.^{1,27,43} Faculties can support individuals involved in assessment and validation by allocating resources to support the appropriate administration of an assessment.

There are potentially more human-associated factors that influence the administration of an assessment. These influences include assessor or learner factors such as anxiety, lack of knowledge/comfort about the assessment strategy²⁹ or misunderstanding of the goals of the assessment by any member included in the assessment process.⁴⁴ For instance, if students feel that their expectations of an assessment are not being met they may not engage with an assessment as intended, undermining the intended interpretations of scores.⁴⁴ As mentioned above, transparency of the assessment process is indispensable and the purpose

of the assessment and the intended use of scores must be known by assessors and learners. Assessors must be supported in terms of understanding the assessment purpose and intended score use, and depending on the assessment strategy, preparations could include rater training to ensure understanding and/or mastery of the assessment strategy.^{1,10,33,43,45}

Documenting validity evidences according to a structured and flexible plan

A socially responsible approach to validity includes making the validation process explicit, purposeful, comprehensive, flexible and accessible for all individuals involved in the assessment process. These validation processes should be planned in conjunction with planning an assessment or program of assessment, and thus prior to the administration of the assessment and the interpretation of assessment data. If necessary, the validation process could be adapted according to the data collected.

Identify resources available for validation process

Identifying available resources for validation³⁸ early in the assessment and validation planning process will contribute to a purposeful, comprehensive, and socially responsible validation approach within the constraints of available resources. Financial or human resources (e.g. access to psychometricians, peers, or expertise), available time, technical support (human or automatic scoring) will influence decisions about appropriate targets within a validation process,²⁹ and therefore likely influence when, how, and if an assessment will be adopted in a local environment in a way that supports defensible score interpretation. This consideration for the quality of the assessment data generated from an assessment process does require time, incurs costs, and necessitates the contribution of energy and resources. Ellaway, Topps, Pusic⁴² suggest that individuals involved in medical education should consider the social contract "balancing what we give up (time, resources, autonomy, access, shared edges) versus what we get in return (opportunity, intelligence, efficiency, transparency, accountability, and risk)".^(42, p.35) Therefore, while planning the interpretation of the assessment scores, individuals involved in the assessment and validation processes must keep in mind the resources available for a validation process in addition to the purpose and stakes of the assessment.

Strategically choosing validity evidences

Choosing validity evidences to document during a validation process must take into consideration the purpose and stakes of assessment as well as the available resources. Individuals responsible for validation processes should strive to find quality of the evidence supporting score interpretation within the resources available,^{28,38,46} with consideration for the type of assessment strategies (e.g. OSCE, portfolio, written exams, workplace-based assessment). Several resources are available to educators and administrators to help choose and prioritize relevant evidences of validity. Downing,¹⁸ for example, provides explicit options for each of the five evidences of validity included in the Standards⁴⁷ that are relevant to medical or health professions education.

The effort to find the best (and most realistic) way to ensure the quality of assessment often requires going beyond psychometrics^{37,48} as the sole measure of quality. We do not suggest that we abandon psychometric analyses – that would be akin to throwing the baby out with the bathwater. Rather, we encourage the use of psychometrics when and where appropriate, however, they should not be the only marker of quality. We suggest that psychometric analyses should be one among many kinds of evidence of quality, and the relative reliance on psychometric measures of assessment quality may differ depending on the assessment approach and purpose. As an example, we can imagine that it would be inadvisable to rely solely on psychometric measures of quality, such as Cronbach alpha, to document the quality of a portfolio. Validity evidence may be quantitative (e.g. psychometric or statistical analysis) qualitative (e.g. triangulation of data sources) in nature, or a combination of both.^{17,46,49} Throughout the assessment development and validation process, qualitative and quantitative data should be collected and attended to, with a plan as to how best to integrate validity evidence relying on qualitative and quantitative data sources.

Adjusting the validation process according to the documented evidences of validity

Good planning allows individuals that are responsible for the assessments to choose appropriate validity evidences considering the resources available and the current context of the assessment. However, it is difficult to predict everything. Therefore, it is essential to

adapt the planned validation process to accommodate unforeseen circumstances. For example, unexpected consequences may arise from the interpretation of learning assessment results will need to be documented through the addition of an evidence.

Planning the interpretation of validity evidences and assessment scores

To be coherent with the social imperative of validity, interpretation of validity evidences and assessment scores must be planned, with considerations for the social consequences of assessment and the individual and societal levels. In other words, before assessment scores are interpreted, and judgments or decisions are made based on those scores, we must ensure that we have collected evidence that supports the quality of the assessment scores to the best of our ability. We must have data that suggests the assessment data was collected defensibly, grounded in well-designed assessment tools and systems, generated through transparent processes, and that the interpretations of scores are fair. We must do the best we can, considering the stakes and the purpose of assessment, to ensure that we have evidence supporting our score interpretation before sharing those with the learner, program, and educational leadership. Ensuring the strength of assessment development and validation practices prior to sharing score interpretation and use - which can have consequences for learners and for society – enables defensible score interpretation and facilitates the mitigation of potential unintended consequences.

Examine quality of scores collected given the intended score use

Decisions based on the interpretation of assessment data yield actions (e.g. pass or fail; continue in practice or not; identification of areas for improvement),^{1,27,46} and these actions should align with the assessment’s purpose and stakes (e.g. low stake such as feedback on the progression of the learner; high stake such as certification, licensure or graduation). Individual involved in the validation process must plan how administrators and educators will use the scores to support decisions and actions, or how the scores will be combined and interpreted. Questions such as “Does the score involved a pass or fail decision?” or “What does the assessors do when scores are diverging from different assessors or various assessments?” must be determined before the collection of the scores to ensure that

assessment purpose, the stakes of assessment, and the approaches to validation all support the score interpretation and use.

Critically review the quality of evidence documented through the validation process

Throughout these guideposts, we have strongly encouraged the use of appropriate and transparent documentation of validity evidence obtained during the entire assessment and validation process (development, administration and monitoring). The transparent documenting of decisions, decision points, and evidence collected add credibility and defensibility of the assessment. However, this data should be continually scrutinized and cyclically collected, as validity is not a property of a given assessment tool, program, or even assessment administration. Considering the documentation of a given assessment or program of assessment as a ‘living document’ that is open to being refined, critiqued, and adapted situates validation within a continual quality improvement process. Additionally, open and transparent documentation regarding the quality monitoring of individual assessments and assessment programs makes it possible to publicly defend – to the learner and the society- the decisions resulting from the assessment scores.³⁵

Conclusion

We aimed to make the concept of validity as a social imperative accessible to individuals responsible for assessments and those enacting validation processes. This description of considerations for how to apply the concept of validity as a social imperative will support those responsible for the planning, administering, and monitoring of assessments in collecting and documenting evidence of validity that could be perceived as credible by society. Transparent and deliberate documentation of assessment and validation practices may decrease the potential negative or unintended consequences of assessment on individual learners and society as a whole. Using rigorous methods, we attempted to elaborate guideposts in what we hope is a non-prescriptive manner.^{8,9} We have intentionally suggested areas for consideration and planning, rather than a prescriptive list of necessary practices aligned with the concept of validity as a social imperative. We hope that this encourages those working with assessments to reflect on their assessment development and validation practices in order to generate data that better supports

decisions dependent on high quality assessment scores, while considering individual context and available resources

Acknowledgments: None.

Funding: This study was supported by the *MEES-Universités programme* (in the form of a scholarship obtained by MM), the Social Sciences and Humanities Research Council (in the form of a scholarship provided by CSO and MY from grant no. 435- 2014-2159) and the *Paul Grand'Maison de la Société des médecins* Research Chair in Medical Education (in the form of a scholarship provided by CSO).

Disclaimers: None.

Ethical approval: The ethical review board of Education and Social Sciences, Université de Sherbrooke, approved this study on June 22, 2016 (file number: 2016-34-ESS).

References

1. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment and Kane's validity perspective. *Med Educ*. 2012;46(1):38-48. doi:10.1111/j.1365-2923.2011.04098.x
2. Schuwirth LWT, van der Vleuten CPM. Programmatic assessment: From assessment of learning to assessment for learning. *Med Teach*. 2011;33(6):478-485. doi:10.3109/0142159X.2011.565828
3. Grant J. Principles of Curriculum Design. In: Swanwick T, Forrest K, O'Brien BC, eds. *Understanding Medical Education: Evidence, Theory and Practice*. 3rd ed. John Wiley & Sons; 2019:71-88.
4. Darling-Hammond L, Snyder J. Authentic assessment of teaching in context. *Teach Teach Educ*. 2000;16(5-6):523-545. doi:10.1016/S0742-051X(00)00015-9
5. Govaerts M, van der Vleuten CPM. Validity in work-based assessment: Expanding our horizons. *Med Educ*. 2013;47(12):1164-1174. doi:10.1111/medu.12289
6. Lucey CR, Thibault GE, Ten Cate O. Competency-Based, Time-Variable Education in the Health Professions: Crossroads. *Acad Med J Assoc Am Med Coll*. 2018;93:S1-S5. doi:10.1097/ACM.0000000000002080
7. Carraccio C, Englander R, Holmboe ES, Kogan JR. Driving Care Quality: Aligning Trainee Assessment and Supervision Through Practical Application of Entrustable Professional Activities, Competencies, and Milestones. *Acad Med J Assoc Am Med Coll*. 2016;91(2):199-203. doi:10.1097/ACM.0000000000000985
8. Kane MT. Validation. In: Brennan RL, editor. *Educational Measurement*. 4th ed. Wesport, CT: American Council on Education/Praeger; 2006:17-64.
9. Messick S. Standards of Validity and the Validity of Standards in Performance Assessment. *Educ Meas Issues Pract*. 1995;14(4):5-8. doi:10.1111/j.1745-3992.1995.tb00881.x
10. Berendonk C, Stalmeijer RE, Schuwirth LWT. Expertise in performance assessment: Assessors' perspectives. *Adv Heal Sci Educ*. 2013;18(4):559-571. doi:10.1007/s10459-012-9392-x
11. Hodges B. Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*. 2013;35(7):564-568. doi:10.3109/0142159x.2013.789134
12. Gingerich A, Kogan J, Yeates P, Govaerts M, Holmboe E. Seeing the "black box" differently: Assessor cognition from three research perspectives. *Med Educ*. 2014;48(11):1055-1068. doi:10.1111/medu.12546
13. Ginsburg S, McIlroy J, Oulanova O, Eva K, Regehr G. Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Acad Med*. 2010;85(5):780-786. doi:10.1097/ACM.0b013e3181d73fb6
14. Schuwirth LWT, van der Vleuten CPM. A plea for new psychometric models in educational assessment. *Med Educ*. 2006;40(4):296-300. doi:10.1111/j.1365-2929.2006.02405.x

15. St-Onge C, Young M, Eva KW, Hodges B. Validity: One word with a plurality of meanings. *Adv Heal Sci Educ*. 2017;22(4):853-867. doi:doi:10.1007/s10459-016-9716-3
16. Cook DA, Beckman TJ. Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *Am J Med*. 2006;119(2):166.e7-166.e16. doi:10.1016/j.amjmed.2005.10.036
17. Cook DA, Brydges R, Ginsburg S, Hatala R. A contemporary approach to validity arguments: A practical guide to Kane's framework. *Med Educ*. 2015;49(6):560-575. doi:10.1111/medu.12678
18. Downing SM. Validity: On the meaningful interpretation of assessment data. *Med Educ*. 2003;37(9):830-837. doi:10.1046/j.1365-2923.2003.01594.x
19. Cizek GJ, Rosenberg SL, Koons HH. Sources of Validity Evidence for Educational and Psychological Tests. *Educ Psychol Meas*. 2008;68(3):397-412. doi:10.1177/0013164407310130
20. Cizek GJ, Bowen D, Church K. Sources of validity evidence for educational and psychological tests: A follow-up study. *Educ Psychol Meas*. 2010;70(5):732-743. doi:10.1177/0013164410379323
21. Cook DA, Zendejas B, Hamstra SJ, Hatala R, Brydges R. What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Adv Heal Sci Educ*. 2014;19(2):233-250. doi:10.1007/s10459-013-9458-4
22. Cook DA, Brydges R, Zendejas B, Hamstra SJ, Hatala R. Technology-Enhanced Simulation to Assess Health Professionals. *Acad Med*. 2013;88(6):872-883. doi:10.1097/ACM.0b013e31828ffdcf
23. Marceau M, Gallagher F, Young M, St-Onge C. Validity as a social imperative for assessment in health professions education: a concept analysis. *Med Educ*. 2018;52(6):641-653. doi:10.1111/medu.13574
24. Marceau M, St-Onge C, Gallagher F, Young M. Validity as a social imperative: Users' and leaders' perceived acceptability and anticipated feasibility. 2019. Unpublished.
25. Thorne S. *Interpretative Description: Qualitative Research for Applied Practice*. New York: Routledge; 2016.
26. Miles MB, Huberman AM, Saldaña J. *Qualitative Data Analysis: A Methods Sourcebook*. 3rd ed. Thousand Oaks, CA; Sage; 2014.
27. van der Vleuten CPM, Schuwirth LWT, Scheele F, Driessen EW, Hodges B. The assessment of professional competence: building blocks for theory development. *Best Pract Res Clin Obstet Gynaecol*. 2010;24(6):703-719. doi:10.1016/j.bpobgyn.2010.04.001
28. van der Vleuten CPM, Schuwirth LWT, Driessen EW, et al. A model for programmatic assessment fit for purpose. *Med Teach*. 2012;34(3):205-214. doi:10.3109/0142159X.2012.652239
29. Baartman L, Gulikers J, Dijkstra A. Factors influencing assessment quality in higher vocational education. *Assess Eval High Educ*. 2013;38(8):978-997. doi:10.1080/02602938.2013.771133

30. Dijkstra J, van der Vleuten CPM, Schuwirth LWT. A new framework for designing programmes of assessment. *Adv Heal Sci Educ.* 2010;15(3):379-393. doi:10.1007/s10459-009-9205-z
31. Timmerman AA, Dijkstra J. A practical approach to programmatic assessment design. *Adv Heal Sci Educ.* 2017;22(5):1169-1182. doi:10.1007/s10459-017-9756-3
32. Crossley J, Humphris G, Jolly B. Assessing health professionals. *Med Educ.* 2002;36(9):800-804. doi:10.1046/j.1365-2923.2002.01294.x
33. Shumway JM, Harden RM. AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Med Teach.* 2003;25(6):569-584. doi:10.1080/0142159032000151907
34. Kasanda CD, Mitonga KH, Veii K, Zimba RF. Medical and pharmacy students' perceptions of the grading and assessment practices. *Front Psychol.* 2013;4(July):423. doi:10.3389/fpsyg.2013.00423
35. Prideaux D, Roberts C, Eva K, et al. Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* 2011;33(3):215-223. doi:10.3109/0142159X.2011.551560
36. Schuwirth LWT, van der Vleuten CPM. How to Design a Useful Test: The Principles of Assessment. In: Swanwick T, Forrest K, O'Brien BC, eds. *Understanding Medical Education: Evidence, Theory and Practice.* 3rd ed. Wiley Blackwell; 2019:277-289.
37. Eva KW, Bordage G, Campbell C, et al. Towards a program of assessment for health professionals: from training into practice. *Adv Heal Sci Educ.* 2015;21(4):1-17. doi:10.1007/s10459-015-9653-6
38. van der Vleuten, Schuwirth LWT. Assessing professional competence: from methods to programmes. *Med Educ.* 2005;39(3):309-317. doi:10.1111/j.1365-2929.2005.02094.x
39. Schuwirth L, Colliver J, Gruppen L, et al. Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach.* 2011;33(3):224-233. <http://10.0.12.37/0142159X.2011.551558>.
40. Fotheringham D. Triangulation for the assessment of clinical nursing skills: A review of theory, use and methodology. *Int J Nurs Stud.* 2010;47(3):386-391. doi:10.1016/j.ijnurstu.2009.09.004
41. Cook DA, Kuper A, Hatala R, Ginsburg S. When assessment data are words: Validity evidence for qualitative educational assessments. *Acad Med.* 2016;91(10):1359-1369. doi:10.1097/ACM.0000000000001175
42. Ellaway RH, Topps D, Pusic M. Data, Big and Small: Emerging Challenges to Medical Education Scholarship. 2019;94(1):31-36. doi:10.1097/ACM.0000000000002465
43. Roberts C, Newble D, Jolly B, Reed M, Hampton K. Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Med Teach.* 2006;28(6):535-543. doi:10.1080/01421590600711187

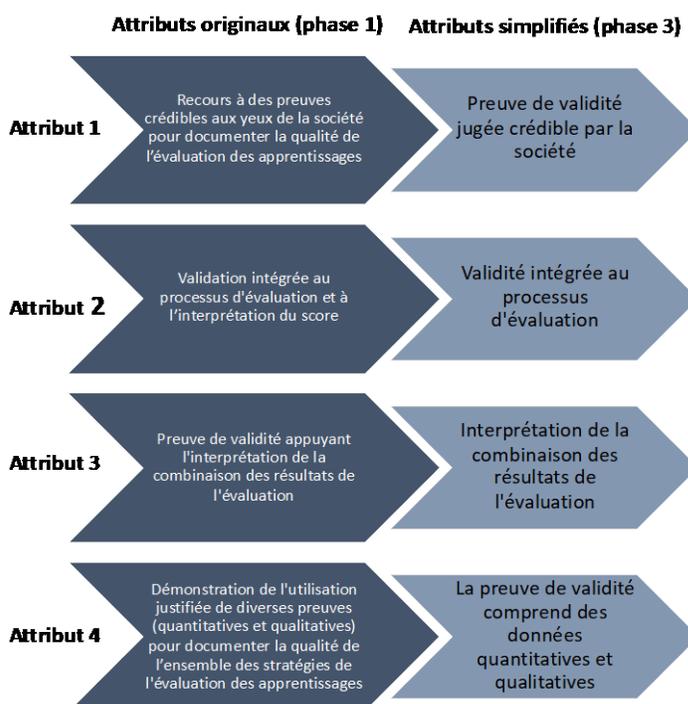
44. Ricci, M., St-Onge, C., Xiao, J., & Young, M. (2018). Students as stakeholders in assessment: how students perceive the value of an assessment. *Perspectives on Medical Education*, 7(6), 352–361. <https://doi.org/10.1007/s40037-018-0480-3>
45. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: The do's, don'ts and don't knows of direct observation of clinical skills in medical education. *Perspect Med Educ*. 2017;6(5):286-305. doi:10.1007/s40037-017-0376-7
46. Cook DA, Lineberry M. Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Acad Med*. 2016;91(6):785-795. doi:10.1097/ACM.0000000000001114
47. AERA, APA, NCME. Validity. In: *Standards for Educational and Psychological Testing*; 2014:11-31.
48. Crossley J. Validity and truth in assessment. *Med Educ*. 2013;47(12):1152-1154. doi:10.1111/medu.12317
49. DeLuca C. Interpretive validity theory: mapping a methodology for validating educational assessments. *Educ Res*. 2011;53(3):303-320. doi:10.1080/00131881.2011.598659

6.2 Résultats complémentaires

Nous décrivons dans cette section l'évolution du concept de validité en tant qu'impératif social au cours des trois phases de la recherche. Le processus itératif menant aux résultats présentés dans l'article 3 a incité l'équipe de recherche à 1) simplifier des attributs, 2) préciser le terme « société », 3) raffiner la description des attributs et 4) réorganiser l'ordre des attributs.

Premièrement, nous avons simplifié la formulation de chacun des attributs dans l'objectif de les rendre plus compréhensibles (figure 5). Dans la colonne de gauche, l'attribut original dénombre plus de mots. Dans la colonne de droite, la formulation de l'attribut est plus concise.

Figure 5
Simplification des attributs



Deuxièmement, comme les questionnements des participants portaient principalement sur les termes « social » et « société », nous avons précisé notre définition de société. L'étudiante s'est inspirée de la définition proposée par Cook & Lineberry (2016) pour qui

la société est constituée « d'apprenants, d'enseignants et d'établissements d'enseignement; de patients, de professionnels de la santé et d'établissements de santé; et même de la société en général ». (Traduction libre de Cook & Lineberry, 2016, p. 788). Cette clarification facilitera, entre autres, la compréhension du lecteur quant à l'attribut 1 (preuve de validité jugée crédible par la société) ou encore la description de l'attribut 2 (prise en compte des conséquences potentielles que l'interprétation des résultats d'évaluation pourrait avoir sur la personne, l'établissement d'enseignement et la société).

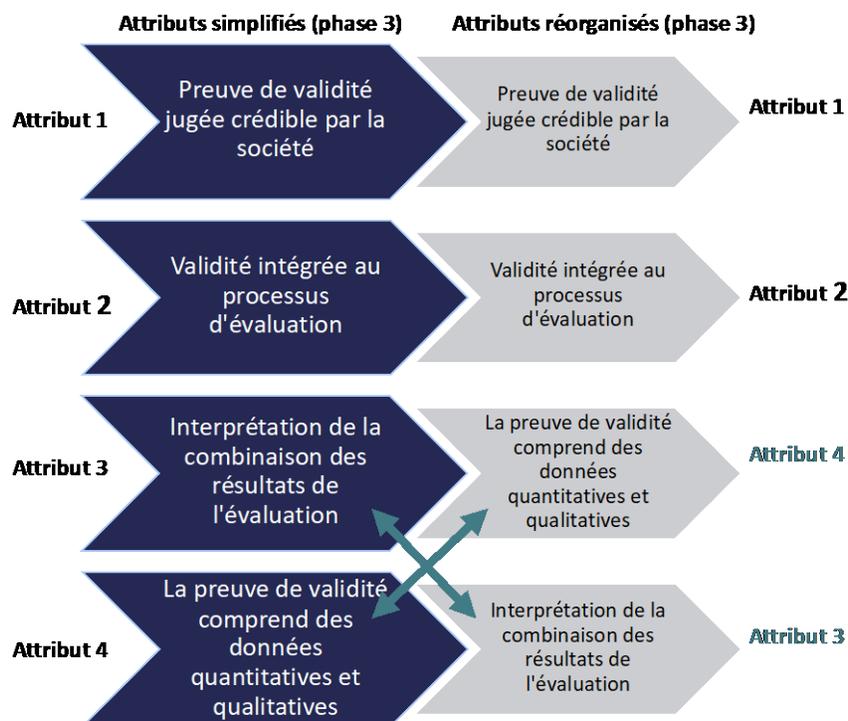
Troisièmement, nous avons raffiné la description des attributs. Pendant les entrevues, les participants avaient émis des réserves sur l'acceptabilité du concept de validité en tant qu'impératif social en raison de l'incompréhension de certains attributs. Rappelons que les parties prenantes avaient reçu une synthèse de l'analyse de concept dans un court document de deux pages (annexes 10 et 18). Nous détaillons dans le tableau 13 les changements apportés aux attributs pendant les trois phases de la recherche.

Tableau 13
Évolution du concept de validité en tant qu'impératif social

Description du concept de validité en tant qu'impératif social (phases 1 et 2) Adapté de l'analyse de concept (Marceau et al., 2018) et du document remis aux participants (annexes 10 et 18)		Description du concept de validité en tant qu'impératif social simplifiée (phase 3)	
Attribut	Description	Attribut	Description
Recours à des preuves crédibles aux yeux de la société pour documenter la qualité de l'évaluation des apprentissages	Les différentes instances professionnelles (établissements d'enseignement et ordres professionnels) doivent pouvoir documenter avec certitude - pour la société - les décisions prises quant au cheminement des apprenants et à leur niveau d'aptitudes à débiter une carrière professionnelle de façon autonome et compétente.	Preuve de validité jugée crédible par la société	Les organismes professionnels (établissements d'enseignement et ordres professionnels) doivent être en mesure de documenter, de manière jugée crédible par la société, les décisions prises concernant le cheminement scolaire des apprenants et leur niveau de compétence à débiter une carrière professionnelle de façon autonome.
Validation intégrée au processus d'évaluation et à l'interprétation du score	La justification des décisions prises durant le processus d'élaboration, d'administration et d'interprétation des résultats repose sur la considération des conséquences potentielles que l'interprétation des résultats d'évaluation pourrait entraîner sur l'individu, l'institution et la société. En plus de mesurer les conséquences découlant de l'interprétation des résultats de l'évaluation des apprentissages, la personne responsable des évaluations doit anticiper les conséquences potentielles liées à cette interprétation et mettre en place des moyens ou des stratégies pour les minimiser. Les conséquences mesurées ne se limitent pas seulement aux impacts liés au construit, mais concernent l'ensemble des conséquences possibles.	Validation intégrée au processus d'évaluation	La preuve de validité inclut la justification des décisions prises lors de l'élaboration et de l'administration d'une évaluation, ainsi que l'interprétation des résultats de l'évaluation. Ces preuves incluent la prise en compte des conséquences potentielles que l'interprétation des résultats d'évaluation pourrait avoir sur la personne, l'établissement d'enseignement et la société.
Preuve de validité appuyant l'interprétation de la combinaison des résultats de l'évaluation	La planification et l'élaboration de l'évaluation (p. ex. : choix des stratégies, objectifs, fréquences, contenus), l'administration ainsi que l'interprétation des résultats de l'évaluation des apprenants doivent être réalisées de façon à considérer un « tout » (le programme d'évaluation dans son entier) qui est supérieur à la somme de ses parties.	Interprétation de la combinaison des résultats de l'évaluation	Les données d'évaluation générées dans le cadre d'un programme d'évaluation sont souvent combinées pour permettre un jugement final. Il convient de mettre davantage l'accent sur les preuves de validité recueillies pour l'interprétation globale des données d'évaluation, plutôt que sur les outils individuels ou les épisodes d'évaluation lorsque les données d'évaluation globales éclairent les décisions relatives aux apprenants.
Démonstration d'une utilisation justifiée de preuves mixtes (quantitatives et qualitatives) pour documenter la qualité de l'ensemble des stratégies de l'évaluation des apprentissages.	Comme les stratégies quantitatives traditionnelles (p. ex. : Alpha de Cronbach) ne sont pas appropriées pour démontrer la qualité d'un ensemble de méthodes d'évaluation des apprentissages (p. ex. : programme d'évaluation), la combinaison de preuves de types quantitatives et qualitatives semble être une option intéressante.	Intégration des données quantitatives et qualitatives	Les preuves de validité doivent être collectées à l'aide de stratégies de validation rigoureuses. Une attention particulière doit être portée aux sources de données quantitatives et qualitatives en tant que preuves de validité légitimes.

Finally, we have reorganized the order of attributes in order to make the elements to be considered during the application of the concept of validity as a social imperative (article 3). Figure 6 allows us to visualize the displacement of attributes 3 and 4.

Figure 6
Organisation des attributs



7. DISCUSSION GÉNÉRALE

Dans une approche par compétences (APC), les personnes responsables de l'évaluation des apprentissages recueillent des données sur la compétence des futurs professionnels par différentes méthodes d'évaluation. Les données sont collectées dans le but de prendre une décision quant à l'admission dans un programme, au rendement et à la progression de l'apprenant ou encore lors de la certification. Par conséquent, l'évaluation des apprentissages entraîne des conséquences sur l'apprenant et sur la société. L'examen de la qualité de l'évaluation des apprentissages s'avère essentiel dans l'optique de s'assurer de la justesse de l'interprétation des résultats (AERA, APA, & NCME, 1999; Messick, 1989; Sireci & Sukin, 2013). Le modèle psychométrique actuel a joué et joue toujours un rôle majeur dans la mesure de la qualité de l'évaluation des apprentissages. Toutefois, il est de plus en plus difficile d'appliquer ce modèle aux nouvelles pratiques évaluatives qui évoluent constamment (Schuwirth & van der Vleuten, 2006; AERA et al., 1999). Pour combler l'écart entre les pratiques évaluatives et les processus de validation existants, des chercheurs en pédagogie des sciences de la santé (p. ex. Gingerich, Kogan, et al., 2014; Gingerich, van der Vleuten, et al., 2014; Hawkins et al., 2015; Hodges, 2013; Holmboe et al., 2017; Schuwirth & Ash, 2013; Schuwirth & van der Vleuten, 2011; van der Vleuten & Schuwirth, 2005; Van Tartwijk & Driessen, 2009) s'interrogent sur l'amélioration des processus de validation par le recours à des preuves de validation qualitatives pour compléter les preuves quantitatives de validation, une vision holistique et englobante de la validité et la considération des conséquences de l'évaluation sur l'apprenant et sur la société. Ces considérations ont été identifiées dans une conceptualisation émergente de la validité : la validité en tant qu'impératif social (St-Onge et al., 2017).

La présente recherche avait pour but d'explorer le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé. Plus spécifiquement, nous voulions 1) décrire le concept émergent de la validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages; 2) explorer l'acceptabilité et la faisabilité anticipée du concept de validité en tant qu'impératif social

et 3) décrire les éléments à considérer lors de l'application du concept de validité en tant qu'impératif social.

Précédemment, nous avons discuté chacun des objectifs à l'intérieur des articles 1, 2 et 3. Dans le chapitre 7, nous souhaitons aller plus loin dans nos réflexions et présenter une discussion sur l'ensemble de la recherche. Nous débuterons par la présentation des principaux constats découlant de la recherche et les discuterons. Par la suite, nous présenterons les forces et les limites au plan méthodologique et ainsi que les retombées pour la pédagogie en sciences de la santé et pour la recherche.

7.1 Principaux constats

Nous relevons cinq principaux constats à l'issue de la démarche de recherche, lesquels sont : 1) l'acceptabilité du concept de la validité en tant qu'impératif social, 2) le lien entre l'impératif social et la responsabilité sociale; 3) la réponse aux préoccupations concernant les limites de processus de validation courants, 4) le recours aux critères de rigueur des devis mixtes et 5) l'imputabilité du processus de validation.

7.1.1 Acceptabilité du concept de validité en tant qu'impératif social

Vérifier l'acceptabilité du concept de validité en tant qu'impératif social représentait une étape incontournable en prévision des futurs développements de ce concept. En ce sens, Sidani et Braden (2011) soulignent l'importance de vérifier l'acceptabilité d'une « intervention » auprès des différentes parties prenantes avant de la mettre en application. Cette précaution n'est d'ailleurs pas nouvelle en pédagogie des sciences de la santé, différents auteurs ayant vérifié l'acceptabilité d'un modèle conceptuel ou encore de méthodes d'évaluation (Dory, Gagnon, De Foy, Duyver, & Leconte, 2010; Massie & Ali, 2016; Murphy et al., 2008; Schultz, Griffiths, & Lacasse, 2015; van der Vleuten & Heeneman, 2016; van der Vleuten et al., 2005).

Les participants (phase 2) sont d'avis que le concept de validité en tant qu'impératif social est pertinent et acceptable, malgré que certains aient observé des similitudes entre cette conceptualisation et les théories modernes de la validité. Il importe de souligner ici la

complémentarité entre le concept de validité en tant qu'impératif social et les théories modernes de la validité. Par le concept de validité en tant qu'impératif social, nous rendons explicites les considérations sociales liées à la validité qui avaient été mises de l'avant de façon implicite par différents théoriciens de la validité, tels que Mislevy (2007), Messick (1989, 1995) et Kane (2006), mais qui ne semblent pas avoir été importées en pédagogie des sciences de la santé.

7.1.2 Lien entre l'impératif social et la responsabilité sociale

Nous avons délibérément sélectionné le terme « impératif social » pour identifier notre concept de validité. Nous souhaitons ainsi distinguer le concept de validité en tant qu'impératif social et celui de la responsabilité sociale. Bien qu'il s'agisse de deux concepts distincts, il est permis de croire que certains éléments qui les caractérisent sont interreliés.

Ces dernières années, le concept de responsabilité sociale a été popularisé et intégré dans les programmes professionnalisants en sciences de la santé. La responsabilité sociale se définit par : *“the obligation to direct their education, research and service activities towards addressing the priority health concerns of the community, region, and/or nation they have a mandate to serve.* (Boelen et al., 1995, p. 3)”. Les responsables de l'élaboration des programmes de formation et les enseignants s'assurent de former des professionnels de la santé qui puissent répondre aux besoins de la population (Boelen, 2011; Boelen et al., 1995; Frenk et al., 2010; Gallagher et al., 2015; Gotzmann, De Champlain, Touchie, & Boulet, 2013; Larkins et al., 2013; Royal College of Physicians and Surgeons of Canada, 2014, 2019a; Tekian, Hodges, Roberts, Schuwirth, & Norcini, 2015; Touchie, 2013; Wen, Greysen, Keszthelyi, Bracero, & de Roos, 2011). De manière plus concrète, la notion de responsabilité sociale influence les décideurs lors de la planification, l'administration et l'évaluation des programmes en sciences de la santé (Boelen, 2018), que ce soit par le choix du contenu enseigné (p. ex. l'enseignement de problèmes de santé fréquents tels que les maladies chroniques), une préoccupation concernant l'accessibilité aux soins (p. ex. offre de stage en région rurale ou auprès d'une clientèle vulnérable) ou l'utilisation judicieuse

des ressources (p. ex. prescription d'une médication moins coûteuse ou choix d'un examen de laboratoire plus accessible).

L'exploration du concept de validité en tant qu'impératif social se justifie par la volonté de combler l'écart entre les pratiques évaluatives utilisées dans une APC et les processus de validation existants. Une meilleure validation des pratiques évaluatives est indispensable pour mesurer les compétences des futurs professionnels de la santé, et ce, dans l'objectif de minimiser les conséquences sur l'individu et sur la société. Bref, nous considérons que la prise en compte des répercussions sur la société est présente dans le concept de validité en tant qu'impératif social et celui de la responsabilité sociale. Bien que la validation des pratiques évaluatives ne soit pas explicitée dans les écrits scientifiques portant sur la responsabilité sociale, nous ne pouvons pas séparer la formation de professionnels compétents de la qualité des méthodes pour les évaluer. Afin de mieux décrire le lien entre les deux concepts, il serait opportun d'explorer la relation entre le concept de validité en tant qu'impératif social et celui de la responsabilité sociale.

7.1.3 Réponse aux préoccupations concernant les limites de processus de validation courants

Le concept de validité en tant qu'impératif social découle des insatisfactions de différents auteurs sur les processus de validation existants et de leurs réflexions sur « comment » faire autrement (Berendonk et al., 2013; Dijkstra et al., 2010; Gingerich, Kogan, et al., 2014; Ginsburg et al., 2010; Hodges, 2013; Holmboe et al., 2017; Schuwirth & van der Vleuten, 2006; Schuwirth & van der Vleuten, 2011; van der Vleuten et al., 2012). Notre recherche avait pour but d'explorer le concept émergent de la validité en tant qu'impératif social. Pour y arriver, nous avons mené une analyse de concept par la recherche d'écrits scientifiques publiés entre 1995 et 2016, pour ensuite interroger différentes parties prenantes en 2016 et 2017. Or, depuis 2016, nous avons remarqué une augmentation du nombre des écrits sur le sujet, ce qui indique que les personnes impliquées en pédagogie des sciences de la santé sont encore aujourd'hui à la recherche de solutions pour diminuer l'écart entre les pratiques évaluatives et les processus de validation existants. La validation des pratiques évaluatives représente toujours un défi pour les personnes responsables de

l'évaluation (Boyd et al., 2018; Cizek, 2016; Gruppen, Ten Cate, Lingard, Teunissen, & Kogan, 2018; Royal College of Physicians and Surgeons of Canada, 2019a, 2019b; Schuwirth & van der Vleuten, 2019). Le concept de validité en tant qu'impératif social s'inscrit dans la conversation actuelle en pédagogie des sciences de la santé. En effet, nous remarquons un intérêt grandissant pour cette conceptualisation. La publication de l'analyse de concept (Marceau et al., 2018) a d'ailleurs contribué à cette conversation dans la communauté scientifique (Alexander & Cleland, 2018; Frank, Sherbino, & Snell, 2018). Le concept de validité en tant qu'impératif social offre une vision différente et prometteuse de la validité ainsi qu'une piste de réponse aux préoccupations grandissantes au sujet des limites des processus de validation.

7.1.4 Recours aux critères de rigueur des devis mixtes

Dans une perspective de validité en tant qu'impératif social, nous allons plus loin dans les réflexions actuelles par la planification délibérée de la documentation des preuves de validité et de leur interprétation dans un ensemble. Cette planification du processus de validation a pour objectif de minimiser les conséquences négatives et imprévues sur l'apprenant et sur la société. Les réflexions actuelles portent principalement sur l'intégration des résultats découlant de l'évaluation des apprentissages provenant de plusieurs sources (méthodes, moments, contextes et observateurs) (Baartman et al., 2013; Dijkstra et al., 2010; Hodges, 2013; Lockyer et al., 2017; van der Vleuten et al., 2012; van der Vleuten & Schuwirth, 2005). L'utilisation des critères de rigueur reconnus en recherche qualitative (p. ex. la crédibilité ou la transférabilité) a été soulevée comme une solution envisagée pour porter un regard sur l'ensemble des données recueillies sur la performance d'un apprenant (Hodges, 2013; Lockyer et al., 2017; van der Vleuten et al., 2010; Van Tartwijk & Driessen, 2009). Par exemple, afin de juger si l'apprenant est apte à poursuivre son cheminement, un comité pourrait examiner à posteriori les résultats provenant d'un portfolio et d'une évaluation en milieu de travail.

À notre connaissance, peu d'écrits abordent comment poser un jugement sur un ensemble de preuves de validité. L'utilisation des éléments à considérer (phase 3) lors de l'application du concept de validité en tant qu'impératif social exige une réflexion

préalable de la part des personnes responsables de l'évaluation des apprentissages, non seulement sur les décisions concernant l'évaluation des apprentissages, mais également sur le processus de validation. L'intégration de preuves de validation de nature qualitative (p. ex. jugement d'experts, triangulation des sources de données) et quantitative (p. ex. analyses psychométriques) semble être une option pour s'assurer de la qualité des pratiques évaluatives utilisées actuellement en pédagogie des sciences de la santé. Cette combinaison de données quantitatives et qualitatives se rapproche des devis de recherche mixtes qui vont au-delà de la simple collecte de données quantitatives et qualitatives. Les chercheurs qui ont recours aux devis mixtes doivent intégrer et analyser les données dans leur ensemble (Creswell & Creswell, 2018). Les critères de rigueur des devis mixtes incluent également les critères de rigueur quantitatifs pour le volet quantitatif et les critères de rigueur qualitatifs pour le volet qualitatif (Creswell & Creswell, 2018).

Tashakkori & Teddlie (2008) proposent une démarche permettant de vérifier la qualité d'un devis mixte (O'Cathain, 2010). Les aspects de la démarche portent sur 1) la planification de la qualité, 2) la conception de la qualité, 3) la collecte et l'analyse des données, 4) la rigueur de l'interprétation, 5) la transférabilité des inférences, 6) la documentation de la qualité, 7) la synthèse des données et 8) l'utilité de la recherche. L'outil d'évaluation de la qualité méthodologique des études qualitatives, quantitatives et mixtes (*Mixed Methods Appraisal Tool-MMAT*) représente un autre exemple (Hong et al., 2018). Les auteurs de cet outil proposent des critères de qualité méthodologiques spécifiques pour les études de méthodes mixtes comme la justification de l'utilisation des méthodes mixtes, la description de l'intégration des résultats quantitatifs et qualitatifs ainsi que la mention des divergences entre les résultats quantitatifs et qualitatifs (Hong et al., 2018).

Les éléments à considérer pour appliquer le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages (phase 3) sont cohérents avec les critères de rigueur pour les devis mixtes décrits dans ces exemples. L'utilisation des critères de qualité spécifiques aux devis mixtes pourrait s'avérer une solution pour juger de la qualité d'un ensemble de preuves de validité. Les points communs entre les devis mixtes et l'évaluation programmatique appuient cette proposition. Par exemple, dans une

évaluation programmatique, les personnes responsables de l'évaluation sélectionnent des preuves de validité quantitatives (p.ex. analyses psychométriques) pour vérifier la qualité des données de nature quantitative (p. ex. QCM) et des preuves de validité qualitatives (p.ex. triangulation) dans le cas des données de nature qualitative (p. ex. un portfolio). Par la suite, ces personnes doivent documenter de manière détaillée comment elles ont procédé pour intégrer les données afin de poser un jugement sur la qualité des pratiques évaluatives. Ainsi, à l'instar des recherches avec devis mixtes, l'évaluation des apprentissages implique l'interprétation de données qualitatives, quantitatives ainsi que l'intégration de ces données, d'où l'intérêt de considérer les critères utilisés pour juger de la qualité des méthodes de recherche mixtes.

7.1.5 Imputabilité du processus de validation

Nous retrouvons dans les écrits scientifiques différentes stratégies pour réaliser un processus de validation (AERA et al., 2014; Downing, 2003; Kane, 2006; Messick, 1995). À la suite de leur participation au projet de recherche, certains membres du personnel enseignant ont mentionné être davantage conscientisés à leur responsabilité en ce qui concerne le processus de validation.

Plusieurs éléments peuvent expliquer la méconnaissance des participants à propos de leur rôle comme leur conceptualisation de la validité, la formation insuffisante en évaluation des apprentissages ou les limites de l'application des processus de validation. En premier lieu, la conceptualisation de la validité (c.-à-d. conceptualisation classique de la validité ou théories modernes de la validité) de la personne responsable de l'évaluation pourrait influencer ses décisions concernant le processus de validation (St-Onge et al., 2017). Dans la présente recherche, la majorité du personnel enseignant rencontré est composé de professeurs qui enseignent dans les programmes en sciences de la santé. Il est possible que la vision des participants à la recherche corresponde à leur expérience de recherche clinique et à la validation d'outils. En deuxième lieu, comme les participants du profil « personnel enseignant » étaient des cliniciens (médecins, infirmiers et infirmières, ergothérapeute et physiothérapeute), il se peut que ces personnes aient priorisé la participation à des activités de formation continue dans leur domaine clinique plutôt qu'en pédagogie des sciences de

la santé. Finalement, les participants ont partagé ne pas savoir comment effectuer un processus de validation pour vérifier la qualité de leurs méthodes d'évaluation. Selon Cizek (2012), il peut être difficile de documenter l'ensemble des preuves de validité décrites par Messick. La preuve de conséquences et la preuve du processus de réponse sont souvent absentes dans les écrits scientifiques en pédagogie des sciences de la santé (Cook & Lineberry, 2016; Labbé et al., 2018).

Les hypothèses décrites ci-dessus pourraient expliquer la méconnaissance des participants du profil « personnel enseignant » concernant leur rôle dans le processus de validation. Cette prise de conscience pourrait également s'expliquer par la confusion liée à l'imputabilité du processus de validation. Les résultats découlant des entrevues avec différentes parties prenantes démontrent que la responsabilité du processus de validation ne semble pas être clairement établie. Selon leur compréhension, des participants ont nommé que la responsabilité de la qualité de l'évaluation des apprentissages revient au programme ou encore aux organismes accréditeurs.

Pour les auteurs du livre de référence *Standards for Educational and Psychological Testing*, les utilisateurs du test et les développeurs du test sont responsables de la qualité des pratiques évaluatives (AERA et al., 2014). Pour Cizek (2012, 2016b), la justification de l'utilisation d'un test incombe aux utilisateurs du test et aux décideurs (p. ex. instances professionnelles) alors que la responsabilité de la qualité des inférences repose sur les personnes qui élaborent le test. Selon Albanese et al. (2010), la responsabilité de la qualité de l'évaluation des compétences incombe aux facultés. Il est aussi essentiel de se questionner sur l'imputabilité de la preuve de conséquences. La personne qui développe la méthode d'évaluation est-elle responsable de l'emploi qu'en fait son utilisateur? Bien que les développeurs et les utilisateurs aient tous les deux un rôle à jouer dans la validation des pratiques évaluatives, Davies (2012) et Kane (2006, 2013) précisent que les utilisateurs de test sont responsables de leur choix quant à l'utilisation d'un test ou d'une méthode d'évaluation.

Cette responsabilité partagée peut mener à la confusion chez les différentes parties prenantes. Pour mettre en œuvre la validité en tant qu'impératif social, il est primordial d'identifier explicitement à qui incombe la responsabilité du processus de validation. Selon le contexte, il peut s'agir d'une personne ou d'un groupe de personnes comme des membres du personnel enseignant dans un programme en sciences de la santé ou encore des personnes œuvrant auprès d'un organisme accréditeur.

7.2 Forces et limites méthodologiques

En complément de ce qui a été souligné dans les articles, nous décrivons dans cette section les forces et les limites de la recherche exploratoire en trois phases ainsi que les méthodes utilisées pour chacune des phases de la recherche.

7.2.1 Recherche exploratoire en trois phases

Les trois phases de la recherche nous ont permis d'examiner différentes perspectives sur le concept de la validité en tant qu'impératif social. Nous avons planifié chacune des phases pour assurer la faisabilité de l'ensemble de la recherche et la complémentarité des méthodes (O'Cathain, 2010). Les résultats recueillis à chacune des phases ont contribué à la phase suivante. Nous avons effectué une analyse de concept (phase 1) pour définir la validité en tant qu'impératif social à partir des écrits scientifiques. Nous avons ensuite mené des entrevues auprès de deux profils de participants (phase 2) afin de poursuivre l'exploration du concept de la validation de l'évaluation des apprentissages selon un impératif social. Finalement, nous avons intégré les données des phases précédentes (phase 3) pour amener le concept le plus près possible de son application. Le paradigme constructiviste a guidé nos décisions méthodologiques, ce qui a permis d'assurer une cohérence à l'ensemble de la recherche (Creswell & Creswell, 2018; Morse et al., 2006; O'Cathain, 2010).

7.2.2 *Analyse de concept*

Dans la phase 1, nous avons utilisé la méthodologie de Rodgers (Rodgers, 1989; Rodgers & Knafl, 2000) pour mener une analyse de concept. Le choix de la méthode est cohérent avec le cadre paradigmatique constructiviste, la description du concept étant centrée sur les différentes perspectives d'auteurs sur la validité en tant qu'impératif social. L'approche inductive a servi à identifier les antécédents, les attributs et le conséquent du concept. Dans l'analyse de concept, il n'est pas requis de tenir compte de la qualité ou du type des documents, ce qui nous a permis d'effectuer une exploration large des écrits (types et disciplines) (Rodgers, 1989; Rodgers & Knafl, 2000). Pour Rodgers, un concept est évolutif et est lié au contexte (Rodgers & Knafl, 2000; Tofthagen & Fagerstrøm, 2010). La description du concept de la validité en tant qu'impératif social correspond à notre compréhension au moment de la collecte et de l'analyse des données (c.-à-d. en 2016). Depuis la publication de de notre analyse de concept, nous avons remarqué que de nouveaux écrits ont contribué à enrichir les connaissances au sujet du concept (p. ex. Holmboe et al., 2017; Lockyer et al., 2017; Timmerman & Dijkstra, 2017) . La publication de l'analyse de concept en 2018 (Marceau et al., 2018) a également stimulé la réflexion des personnes sur le sujet (Alexander & Cleland, 2018; Frank et al., 2018). Les phases subséquentes de la présente étude ont permis d'en raffiner la description et les attributs. En raison de l'évolution du concept et de l'intérêt grandissant pour la validité en tant qu'impératif social, il est possible qu'une deuxième analyse soit nécessaire dans le futur.

Bien que nous ayons suivi un processus rigoureux de collecte de données, il est possible que nous ayons omis certains documents en raison des mots-clés utilisés. Par conséquent, certains documents, qui auraient pu enrichir nos conclusions, pourraient ne pas avoir été repérés. Soulignons la difficulté à mener l'analyse de concept en raison de la nature émergente de la validité en tant qu'impératif social et de la proximité avec les théories modernes documentées dans les écrits. Lors de l'analyse des documents sélectionnés, nous avons observé que les éléments décrivant le concept pouvaient être ténus. Considérant l'émergence du concept, nous avons inclus toutes les données correspondant aux critères d'inclusion. Par exemple, il pouvait s'agir d'une réflexion de l'auteur sur les limites de la

psychométrie dans l'introduction d'un texte ou encore une simple phrase sur la nécessité de se préoccuper des conséquences sur la société dans la conclusion d'un article.

7.2.3 Approche descriptive interprétative

Dans les phases 2 et 3, nous avons utilisé l'approche de recherche descriptive interprétative. Cette approche nous a permis d'explorer un phénomène dans le but de le rendre applicable dans la pratique. L'approche de Thorne (2016) se veut non prescriptive. L'auteure propose toutefois une approche structurée avec une considération pour les critères de rigueur scientifique. Considérant son développement récent, des auteurs (Berterö, 2015; Hunt, 2009) questionnent les frontières avec les autres méthodologies qualitatives qu'ils qualifient de floues (p. ex. théorisation ancrée). Ainsi, il importe de préciser la question de recherche afin de bien justifier le recours à l'approche descriptive interprétative plutôt que l'utilisation d'une autre approche ou méthode de recherche.

L'approche descriptive interprétative est cohérente avec le cadre paradigmatique constructiviste (Hunt, 2009). Dans la phase 2, le cadre paradigmatique a guidé l'élaboration des guides d'entrevues par l'élaboration de questions ouvertes, permettant le point de vue de différentes parties prenantes. Dans la phase 3, le cadre paradigmatique a permis de revoir sous un autre angle les résultats des phases précédentes, en prenant compte de la multiplicité des réalités (auteurs de l'analyse de concept et participants aux entrevues).

7.2.4 Collecte des données

Nous avons recruté deux profils de participants : 1) le personnel enseignant et les membres de comités d'évaluation des apprentissages ainsi que 2) les personnes engagées dans la recherche au Canada. Étant donné la nature exploratoire du concept émergent, nous ne pouvons pas prétendre avoir atteint une saturation des données. Toutefois, nous avons interrogé des personnes provenant de différents milieux (universités, villes) ayant une expérience variée (disciplines, rôles dans le domaine de l'évaluation, expérience). La variété de points de vue recueillis et la triangulation des méthodes utilisées nous ont permis d'obtenir une richesse et une profondeur des données (Malterud et al., 2016; Morse, 2015; Thorne, 2016). Dans le futur, nous croyons qu'il serait intéressant de poursuivre notre

exploration du concept de validité en tant qu'impératif social auprès d'autres parties prenantes comme des apprenants, des usagers du système de santé, des administrateurs ou encore des personnes travaillant pour des organismes accréditeurs.

L'utilisation de groupes de discussion focalisée a pour avantage de favoriser les interactions entre les participants (Morgan, 2019). Bien que nous ayons regroupé des personnes provenant de différents programmes en sciences de la santé (médecine, sciences infirmières, ergothérapie et physiothérapie), les groupes de discussion étaient plutôt homogènes. En effet, nous avons réuni le personnel enseignant et les membres de comités d'évaluation des apprentissages d'une même université, facilitant les échanges à propos de leur expérience dans le domaine de la validité et de l'évaluation dans un environnement similaire. Les groupes de discussion focalisée ont également des limites, comme celles de créer un contexte artificiel où l'influence du groupe est à considérer et de rendre la logistique difficile (Morgan, 2019). Pour diminuer l'impact du manque de naturel de l'environnement et faciliter le rassemblement des participants, nous avons mené les groupes de discussion focalisée dans les universités des participants et adapté nos entrevues aux disponibilités de gens. Les points de vue recueillis sur le concept de validité en tant qu'impératif social ont été de nature positive et négative, ce qui laisse penser que les personnes ont partagé leur opinion sans être influencées par le groupe ou l'équipe de recherche.

L'entrevue individuelle offre la possibilité d'approfondir les réponses des participants et facilite la logistique de l'entrevue (lieu, moment) (Brinkman & Kvale, 2015; Morgan, 2019). Cette méthode de collecte des données était la plus appropriée pour les entrevues des personnes engagées dans la recherche au Canada. Nous avons pu mener des entrevues avec des personnes partout au Canada. Nous avons également rencontré le personnel enseignant et les membres de comités d'évaluation des apprentissages qui n'étaient pas disponibles pour les groupes de discussion. Les deux formats d'entrevues (en groupe et individuelle) sont complémentaires et permettent d'utiliser les forces de chacune des méthodes (Morgan, 2019) tout en optimisant la composition de l'échantillon

La collecte et l'analyse des données n'ont pas été menées simultanément en raison de certaines limites contextuelles (p.ex. la disponibilité des participants et la distance entre les différentes universités). Toutefois, des ajustements et des améliorations ont été apportés au besoin entre chacune des entrevues par l'étudiante et une directrice. Ainsi, l'écoute des entrevues et la rédaction d'une fiche synthèse ont orienté la poursuite de la collecte des données.

7.2.5 Rigueur scientifique

Nous avons porté une attention particulière à la crédibilité, à la transférabilité et à la fiabilité pour assurer la rigueur scientifique du processus de recherche (Morse, 2015; Frambach, van der Vleuten & Durning, 2013).

La nature exploratoire de la recherche et le paradigme constructiviste ont orienté les travaux de façon à explorer diverses perspectives au sujet de la validité en tant qu'impératif social. Ainsi, la triangulation des données (Morse, 2015; Varpio, Ajjawi, Monrouxe, O'Brien, & Rees, 2017) par l'utilisation de plusieurs méthodes (analyse de concept, entrevues individuelles et groupes de discussion focalisée) et le recours à des sources pertinentes (écrits scientifiques et participants provenant de différents programmes et universités) a contribué à la crédibilité des résultats. Par le fait même, la qualité des résultats en découlant favorise la transférabilité de ceux-ci. Nous avons utilisé une méthode d'échantillonnage par choix raisonné afin d'obtenir une variété de points de vue provenant de différentes parties prenantes. Les données provenant des phases 1 et 2 ont été intégrées et transformées lors de la phase 3. Nous avons tenu compte de la subjectivité pendant les différentes étapes de la recherche (planification de la démarche, collecte de données, analyse des données) en nous positionnant comme chercheuses et en approfondissant la perspective des participants (Thorne, 2016). Pour chacune des phases de la recherche, l'étudiante et une directrice ont coanalysé les données et nous avons eu plusieurs discussions et réflexions au sein de l'équipe de recherche. Tel que recommandé, nous avons vérifié notre interprétation des transcriptions des entrevues avec les personnes engagées dans la recherche au Canada (Frambach, van der Vleuten, & Durning, 2013; Morse, 2015; Thorne, 2016; Varpio et al., 2017). Toutefois, nous n'avons pas fait cette vérification

auprès du personnel enseignant et des membres de comités d'évaluation des apprentissages. Bien que nous n'ayons mené qu'une seule entrevue individuelle ou de groupe auprès de ce profil de participants, nous avons interviewé plusieurs groupes et personnes provenant de plus d'une université.

Pour assurer la fiabilité du processus de recherche, nous avons décrit de manière détaillée chacune des phases de la recherche (Morse, 2015; Thorne, 2016), par exemple, les critères d'inclusion et d'exclusion ainsi que notre stratégie d'échantillonnage (Frambach et al., 2013). Nous avons passé un temps relativement limité sur le terrain (c.-à-d. pour certaines universités, nous n'avons fait qu'un moment de collecte), ce qui n'a toutefois pas nui à la production d'une description riche de la validité en tant qu'impératif social. Ceci a pu être possible grâce à la triangulation des méthodes de recherche, des sources d'information et des méthodes de collecte des données ainsi que par la réalisation d'une analyse avancée des données.

Dans l'optique d'assurer la transférabilité des résultats de la recherche, nous avons détaillé les profils des participants (personnel enseignant, membres de comités d'évaluation des apprentissages et personnes engagées dans la recherche), les universités ciblées par la recherche (McGill, Sherbrooke, Laval et Montréal) ainsi que les programmes en sciences de la santé (réadaptation, médecine, sciences infirmières, éducation). Ces informations permettent de juger des possibilités de transférabilité des résultats à des contextes semblables. L'échantillon ne représente pas tous les programmes en sciences de la santé (p. ex. nutrition, orthophonie, travail social) ou encore les universités qui offrent ces programmes (p. ex. Université du Québec à Trois-Rivières). Notre choix de sites universitaires se justifie par la volonté de rejoindre les quatre disciplines ciblées par l'étude dans une même université. Nous avons recruté des personnes intéressées par le sujet de l'évaluation des apprentissages en pédagogie des sciences de la santé, ce qui a pu mener à un biais de sélection. Toutefois, comme nous l'avons expliqué précédemment, les opinions recueillies étaient à la fois positives et négatives au sujet du concept de la validité en tant qu'impératif social, ce qui laisse présager que les participants ont réellement donné leur

avis sur le sujet. Le biais de désirabilité sociale semble minime considérant cette variation dans les points de vue exprimés.

7.3 Retombées

Les résultats découlant du projet de recherche sont susceptibles d'engendrer des retombées pour la pédagogie des sciences de la santé et pour la recherche. Nous décrivons dans cette section les principales implications de ces retombées.

7.3.1 Retombées pour la pédagogie en sciences de la santé

Les répercussions du projet de recherche sont nombreuses pour la pédagogie des sciences de la santé. Par la participation à ce projet, les participants du groupe personnel enseignant et membres de comités d'évaluation des apprentissages ont exprimé avoir maintenant plus conscience de l'importance de la validité et de ses répercussions sur l'apprenant et sur la société. La diffusion des résultats dans plusieurs événements a contribué à mieux faire connaître la validité en tant qu'impératif social et a contribué à la réflexion portant sur l'amélioration des processus de validation et la qualité des nouvelles pratiques évaluatives. À ce jour, nous avons partagé les résultats dans plusieurs congrès, au niveau provincial (Colloque du groupe de recherche en quête de validité 2019), national (Conférence canadienne sur l'éducation médicale 2015, 2018, 2019) et international (Forum international francophone de pédagogie en sciences de la santé 2016 et 2019; Congrès mondial des infirmières et infirmiers francophones du Secrétariat international des infirmières et infirmiers de l'espace francophone – SIDIIEF 2015).

La publication des résultats de l'analyse de concept (Marceau et al., 2018) dans la revue *Medical Education* en juin 2018 a mené à des réactions dans la communauté de pédagogie des sciences de la santé. Dans un commentaire, Alexander & Cleland (2018) soulignent la pertinence du concept de validité en tant qu'impératif social pour les admissions dans les programmes de médecine. L'article a fait l'objet d'une baladodiffusion KeyLIME PodCast #194 (Key Literature in Medical Education). KeyLIME est une balado hebdomadaire réalisée par le Collège royal des médecins et chirurgiens du Canada qui discute des articles touchant des sujets fondamentaux, novateurs ou qui ont une influence sur la pratique

pédagogique. Suivie par 20 000 personnes, cette balado a une portée internationale. Les animateurs Frank, Sherbino, & Snell (2018) mentionnent : “*This a fantastic paper that opens up a new way of thinking about assessment and lends support to the rationale for CBME [Competency based medical education]*”. Comme indiqué dans le PodCast #194, l’article est actuellement considéré à titre de meilleure publication dans le domaine de l’évaluation des apprentissages selon le Collège Royal des médecins et chirurgiens du Canada.

Nous pouvons envisager que les retombées du projet de recherche se répercutent sur les personnes participantes et non participantes (enseignants, les membres de comités d’évaluation des apprentissages, planificateurs des programmes). Les participants ont reconnu la pertinence du concept de validité en tant qu’impératif social, principalement en ce qui concerne la considération des conséquences découlant de l’évaluation des apprentissages sur l’apprenant et sur la société. L’acceptabilité et l’application future du concept émergent pourraient être influencées positivement par la valeur accordée à la responsabilité sociale dans les programmes de formation en sciences de la santé. De plus, bien qu’il s’agisse d’une première ébauche, nous pensons que les éléments à considérer sont concrets et accessibles pour les personnes responsables de l’évaluation des apprentissages qui pourront sélectionner les éléments pertinents selon leur contexte et l’enjeu lié à l’évaluation.

7.3.2 Retombées pour la recherche

La présente recherche a contribué à l’avancement des connaissances au sujet de la validité. Les résultats apportent une meilleure compréhension de la validité en tant qu’impératif social dans le contexte de l’évaluation des apprentissages, notamment en explicitant le rôle social de la validité dans le contexte de l’évaluation des apprentissages et en élaborant des éléments à considérer lors de l’application du concept. En ce sens, la présente étude contribue à la réflexion sur l’amélioration des processus de validation existants. Nous espérons avoir fourni une piste de solution pour diminuer l’écart entre les pratiques évaluatives et les processus de validation existants. Bien que certains éléments du concept de validité en tant qu’impératif social aient des frontières communes avec les

conceptualisations existantes (p. ex. planification du processus de validation, preuve de conséquences), le concept à l'étude est complémentaire aux théories modernes.

Notre recherche ouvre la voie à d'autres études au sujet de la validité. Tous les acteurs concernés (p. ex. chercheurs, personnel enseignant, organismes accréditeurs) pourraient travailler de concert pour poursuivre l'exploration du concept de validité en tant qu'impératif social. D'ailleurs, nous avons avancé quelques propositions de recherches futures dans la discussion générale. Par exemple, il serait pertinent de réviser le concept de validité en tant qu'impératif social par une deuxième analyse à la lumière des nouveaux écrits scientifiques publiés. Aussi, l'exploration de la perception du concept de validité en tant qu'impératif social du point de vue des apprenants, des utilisateurs du système de santé ou des personnes œuvrant auprès de différents organismes accréditeurs apporterait un éclairage plus complet de son acceptabilité et contribuerait au raffinement du concept. D'autres recherches devraient être envisagées pour documenter l'acceptabilité et la faisabilité des éléments à considérer lors de l'application du concept dans la réalité. Finalement, il serait pertinent d'explorer la mise à l'essai de cette conceptualisation dans le contexte de la pédagogie des sciences de la santé.

CONCLUSION

Nous avons exploré le concept de validité en tant qu'impératif social afin de décrire ses caractéristiques, de documenter la perception qu'en ont les gens qui sont impliqués dans le domaine de l'évaluation des apprentissages et d'élaborer des éléments à considérer pour appliquer le concept.

Le concept de validité en tant qu'impératif social a été identifié dans les écrits pour documenter les insatisfactions des personnes responsables de l'évaluation quant aux limites des processus de validation et pour nourrir les réflexions sur une nouvelle manière de voir la validité. Nous croyons que le concept de validité en tant qu'impératif social est une piste de solution prometteuse pour combler l'écart entre les pratiques évaluatives et le processus de validation actuel. Nous croyons que la psychométrie et les théories modernes de la validité sont toujours pertinentes, mais nous offrons des considérations complémentaires afin de les bonifier.

Nous rendons explicite le rôle social de la validité dans le contexte de l'évaluation des apprentissages en pédagogie des sciences de la santé. En facilitant le travail des personnes responsables de l'évaluation dans les établissements d'enseignement ou dans les instances professionnelles, nous espérons que les résultats de la recherche contribueront à l'amélioration de l'évaluation des apprentissages des futurs professionnels de la santé ainsi qu'à l'avancement de la recherche dans ce domaine et qu'ils permettront de minimiser les conséquences sur l'individu et sur la société.

RÉFÉRENCES

- AERA, APA, & NCME. (1966). *Standards for Educational and Psychological Tests and Manuals*. Washington: American Psychological Association.
- AERA, APA, & NCME. (1999). Validity. In *Standards for Educational and Psychological Testing* (pp. 9–24). Washington: American Educational Research Association.
- AERA, APA, & NCME. (2014). Validity. In *Standards for Educational and Psychological Testing and Standard* (pp. 11–31). Washington: American Educational Research Association.
- Albanese, M. A., Mejicano, G., Anderson, W. M., & Gruppen, L. (2010). Building a competency-based curriculum: the agony and the ecstasy. *Advances in Health Sciences Education, 15*(3), 439–454. <https://doi.org/10.1007/s10459-008-9118-2>
- Alexander, K., & Cleland, J. (2018). Satisfying the hydra: the social imperative in medical school admissions. *Medical Education, 52*(6), 587–589. <https://doi.org/10.1111/medu.13586>
- Arcand, L. L., & Neumann, J. a. (2005). Nursing Competency Assessment Across the Continuum of Care. *Journal of Continuing Education in Nursing, 36*, 247–254.
- Assessment Reform Group. (2002). *Assessment for Learning: 10 principles research-based principles to guide classroom practice*. London: Assessment Reform Group.
- Axley, L. (2008). Competency: a concept analysis. *Nursing Forum, 43*(4), 214–222. <https://doi.org/10.1111/j.1744-6198.2008.00115.x>
- Baartman, L., Gulikers, J., & Dijkstra, A. (2013). Factors influencing assessment quality in higher vocational education. *Assessment & Evaluation in Higher Education, 38*(8), 978–997. <https://doi.org/10.1080/02602938.2013.771133>
- Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. . (2007). Teachers' opinions on quality criteria for Competency Assessment Programs. *Teaching and Teacher Education, 23*(6), 857–867. <https://doi.org/10.1016/j.tate.2006.04.043>
- Baird, J. A., Andrich, D., Hopfenbeck, T. N., & Stobart, G. (2017). Assessment and learning: fields apart? *Assessment in Education: Principles, Policy and Practice, 24*(3), 317–350. <https://doi.org/10.1080/0969594X.2017.1319337>
- Baldwin, M. A. (2008). Concept analysis as a method of inquiry. *Nurse Researcher, 15*(2), 49–58. <https://doi.org/10.7748/nr2008.01.15.2.49.c6329>
- Barton, A. J., Armstrong, G., Preheim, G., Gelmon, S. B., & Andrus, L. C. (2009). A national Delphi to determine developmental progression of quality and safety competencies in nursing education. *Nursing Outlook, 57*(6), 313–322. <https://doi.org/10.1016/j.outlook.2009.08.003>
- Beck, C. (2013). *Routledge international handbook of qualitative nursing research*. New York: Routledge.

- Beckman, T. J., Cook, D. A., & Mandrekar, J. N. (2005). What is the validity evidence for assessments of clinical teaching? *Journal Of General Internal Medicine*, *20*(12), 1159–1164.
- Berendonk, C., Stalmeijer, R. E., & Schuwirth, L. W. T. (2013). Expertise in performance assessment: Assessors' perspectives. *Advances in Health Sciences Education*, *18*(4), 559–571. <https://doi.org/10.1007/s10459-012-9392-x>
- Berterö, C. (2015). Developing qualitative methods: Or “same old wine in a new bottle.” *International Journal of Qualitative Studies on Health and Well-Being*, *10*, 1–2. <https://doi.org/10.3402/qhw.v10.27679>
- Bertrand, R., & Blais, J.-G. (2004). *Modèles de mesure. L'apport de la théorie des réponses aux items*. Sainte-Foy: Les Presses de l'Université du Québec.
- Birt, L., Scott, S., Cavers, D., Campbell, C., & Walter, F. (2016). Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qualitative Health Research*, *26*(13), 1802–1811. <https://doi.org/10.1177/1049732316654870>
- Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., Cunnington, & N., D. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine*, *71*(9), 1002–1007.
- Boelen, C. (2011). Consensus mondial sur la responsabilité sociale des facultés de médecine. *Santé Publique*, *23*(3), 247–250. <https://doi.org/10.4414/saez.2011.15893>
- Boelen, C. (2018). Coordinating medical education and health care systems: the power of the social accountability approach. *Medical Education*, *52*(1), 96–102. <https://doi.org/10.1111/medu.13394>
- Boelen, C., Heck, J. E., & World Health Organization. (1995). *Defining and measuring the social accountability of medical schools*. Geneva.
- Bok, H. G. J., Teunissen, P. W., Favier, R. P., Rietbroek, N. J., Theyse, L. F. H., Brommer, H., ... Jaarsma, D. A. D. C. (2013). Programmatic assessment of competency-based workplace learning: when theory meets practice. *BMC Medical Education*, *13*(123), 1–9. <https://doi.org/10.1186/1472-6920-13-123>
- Boley, P., & Whitney, K. (2003). Grade disputes: considerations for nursing faculty. *The Journal of Nursing Education*, *42*(5), 198–203.
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, *111*(4), 1061–1071. <https://doi.org/10.1037/0033-295X.111.4.1061>
- Bouldin, A., & Wilkin, N. (2000). Programmatic assessment in US schools and colleges of pharmacy: A snapshot. *American Journal of Pharmaceutical Education*, *64*(4), 380–387. Retrieved from <http://archive.ajpe.org/legacy/pdfs/aj640406.pdf>
- Boyd, V. A., Whitehead, C. R., Thille, P., Ginsburg, S., Brydges, R., & Kuper, A. (2018). Competency-based medical education: the discourse of infallibility. *Medical Education*, *52*(1), 45–57. <https://doi.org/10.1111/medu.13467>
- Brinkman, S., & Kvale, S. (2015). *InterViews : learning the craft of qualitative research interviewing* (3rd ed.). Los Angeles: SAGE.

- Brown, K. K., Maryman, J., & Collins, T. (2017). An Evaluation of a Competency-Based Public Health Training Program for Public Health Professionals in Kansas. *Journal of Public Health Management & Practice*, 23(5), 447–453. <https://doi.org/10.1097/PHH.0000000000000513>
- Brown, S. (2019). Using assessment and feedback to empower students and enhance their learning. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education: A Handbook for Academic Practitioners* (2nd ed., pp. 50–63). New York: Routledge.
- Callwood, A., Cooke, D., & Allan, H. (2014). Developing and piloting the multiple mini-interview in pre-registration student midwife selection in a UK setting. *Nurse Education Today*, 34(12), 1450–1454. <https://doi.org/10.1016/j.nedt.2014.04.023>
- Carraccio, C., Englander, R., Holmboe, E. S., & Kogan, J. R. (2016). Driving Care Quality: Aligning Trainee Assessment and Supervision Through Practical Application of Entrustable Professional Activities, Competencies, and Milestones. *Academic Medicine: Journal Of The Association Of American Medical Colleges*, 91(2), 199–203. <https://doi.org/10.1097/ACM.0000000000000985>
- Carraccio, C., Englander, R., Van Melle, E., ten Cate, O., Lockyer, J., Chan, M.-K., ... Snell, L. S. (2016). Advancing Competency-Based Medical Education. *Academic Medicine*, 91(5), 645–649. <https://doi.org/10.1097/ACM.0000000000001048>
- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition*, 20(6), 633–642. <https://doi.org/10.3758/BF03202713>
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13. <https://doi.org/10.1111/j.1745-3992.2009.00165.x>
- Cilliers, F. J., Schuwirth, L. W. T., Herman, N., Adendorff, H. J., & van der Vleuten, C. P. M. (2012). A model of the pre-assessment learning effects of summative assessment in medical education. *Advances in Health Sciences Education*, 17(1), 39–53. <https://doi.org/10.1007/s10459-011-9292-5>
- Cilliers, F. J., Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2012). Modelling the pre-assessment learning effects of assessment: evidence in the validity chain. *Medical Education*, 46(11), 1087–1098. <https://doi.org/10.1111/j.1365-2923.2012.04334.x>
- Cizek, G. J., Rosenberg, S. L., & Koons, H. H. (2008). Sources of Validity Evidence for Educational and Psychological Tests. *Educational and Psychological Measurement*, 68(3), 397–412. <https://doi.org/10.1177/0013164407310130>
- Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods*, 17(1), 31–43. <https://doi.org/10.1037/a0026975>
- Cizek, G. J. (2016a). Progress on validity: the glass half full, the work half done. *Assessment in Education: Principles, Policy and Practice*, 23(2), 304–308. <https://doi.org/10.1080/0969594X.2016.1156642>
- Cizek, G. J. (2016b). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy and Practice*, 23(2), 212–225. <https://doi.org/10.1080/0969594X.2015.1063479>

- Cizek, G. J., Bowen, D., & Church, K. (2010). Sources of validity evidence for educational and psychological tests: A follow-up study. *Educational and Psychological Measurement*, *70*(5), 732–743. <https://doi.org/10.1177/0013164410379323>
- Cizek, G. J., Kosh, A. E., & Toutkoushian, E. K. (2018). Gathering and Evaluating Validity Evidence: The Generalized Assessment Alignment Tool. *Journal of Educational Measurement*, *55*(4), 477–512. <https://doi.org/10.1111/jedm.12189>
- Colliver, J. A., Conlee, M. J., & Verhulst, S. J. (2012). From test validity to construct validity ... and back? *Medical Education*, *46*(4), 366–371. <https://doi.org/10.1111/j.1365-2923.2011.04194.x>
- Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education*, *49*(6), 560–575. <https://doi.org/10.1111/medu.12678>
- Cook, D. A., & Lineberry, M. (2016). Consequences Validity Evidence: Evaluating the Impact of Educational Assessments. *Academic Medicine*. <https://doi.org/10.1097/ACM.0000000000001114>
- Cook, D. A., & Beckman, T. J. (2006). Current Concepts in Validity and Reliability for Psychometric Instruments: Theory and Application. *The American Journal of Medicine*, *119*(2), 166.e7-166.e16. <https://doi.org/10.1016/j.amjmed.2005.10.036>
- Cook, D. A., Kuper, A., Hatala, R., & Ginsburg, S. (2016). When assessment data are words: Validity evidence for qualitative educational assessments. *Academic Medicine*, *91*(10), 1359–1369. <https://doi.org/10.1097/ACM.0000000000001175>
- Cook, D. A., Brydges, R., Zendejas, B., Hamstra, S. J., & Hatala, R. (2013). Technology-Enhanced Simulation to Assess Health Professionals. *Academic Medicine*, *88*(6), 872–883. <https://doi.org/10.1097/ACM.0b013e31828ffdcf>
- Cook, D. A., Zendejas, B., Hamstra, S. J., Hatala, R., & Brydges, R. (2014). What counts as validity evidence? Examples and prevalence in a systematic review of simulation-based assessment. *Advances in Health Sciences Education*, *19*(2), 233–250. <https://doi.org/10.1007/s10459-013-9458-4>
- Cooper, N. J. (2000). Facilitating Learning from Formative Feedback in Level 3 Assessment. *Assessment & Evaluation in Higher Education*, *25*(3), 279–291. <https://doi.org/10.1080/713611435>
- Coulby, C., Hennessey, S., Davies, N., & Fuller, R. (2011). The use of mobile technology for work-based assessment: the student experience. *British Journal of Educational Technology*, *42*(2), 251–265. <https://doi.org/10.1111/j.1467-8535.2009.01022.x>
- Creswell, J. W. (2014). *Research design : qualitative, quantitative, and mixed methods approaches* (4th ed.). Thousand Oaks: SAGE.
- Creswell, J. W., & Creswell, J. D. (2018). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (5th ed.). Los Angeles: SAGE.
- Croft, H., Gilligan, C., Rasiah, R., Levett-Jones, T., & Schneider, J. (2019). Current Trends and Opportunities for Competency Assessment in Pharmacy Education—A Literature Review. *Pharmacy*, *7*(2), 67. <https://doi.org/10.3390/pharmacy7020067>

- Crossley, J. (2013). Validity and truth in assessment. *Medical Education*, 47(12), 1152–1154. <https://doi.org/10.1111/medu.12317>
- Crossley, J., Humphris, G., & Jolly, B. (2002). Assessing health professionals. *Medical Education*, 36(9), 800–804. <https://doi.org/10.1046/j.1365-2923.2002.01294.x>
- Crossley, J., & Jolly, B. (2012). Making sense of work-based assessment: ask the right questions, in the right way, about the right things, of the right people. *Medical Education*, 46(1), 28–37. <https://doi.org/10.1111/j.1365-2923.2011.04166.x>
- Curran, V., Hollett, A., Casimiro, L. M., McCarthy, P., Banfield, V., Hall, P., ... Wagner, S. (2011). Development and validation of the interprofessional collaborator assessment rubric ((ICAR)). *Journal of Interprofessional Care*, 25(5), 339–344. <https://doi.org/10.3109/13561820.2011.589542>
- Curran, V. R., Butler, R., Duke, P., Eaton, W. H., Moffatt, S. M., Sherman, G. P., & Pottle, M. (2012). Effectiveness of a Simulated Clinical Examination in the Assessment of the Clinical Competencies of Entry-Level Trainees in a Family Medicine Residency Programme. *Assessment & Evaluation in Higher Education*, 37(1), 99–112. <https://doi.org/10.1080/02602938.2010.515009>
- Dannefer, E. F. (2013). Beyond assessment of learning toward assessment for learning: educating tomorrow's physicians. *Medical Teacher*, 35(7), 560–563. <https://doi.org/10.3109/0142159X.2013.787141>
- Dannefer, E. F., & Henson, L. C. (2007). The portfolio approach to competency-based assessment at the Cleveland Clinic Lerner College of Medicine. *Academic Medicine: Journal Of The Association Of American Medical Colleges*, 82(5), 493–502. Retrieved from <http://hmu.edu.iq/About/Overview.aspx>
- Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education*, 16(5–6), 523–545. [https://doi.org/10.1016/S0742-051X\(00\)00015-9](https://doi.org/10.1016/S0742-051X(00)00015-9)
- Davies, A. (2012). Kane, validity and soundness. *Language Testing*, 29(1), 37–42. <https://doi.org/10.1177/0265532211417213>
- Dedoose. (2018). Web application for managing, analyzing, and presenting qualitative and mixed method research data. Los Angeles, CA: SocioCultural Research Consultants, LLC. Retrieved from www.dedoose.com.
- DeLuca, C. (2011). Interpretive validity theory: mapping a methodology for validating educational assessments. *Educational Research*, 53(3), 303–320. <https://doi.org/10.1080/00131881.2011.598659>
- Denzin, N. K., & Lincoln, Y. S. (2018). *The SAGE handbook of qualitative research* (5th ed.). Los Angeles: SAGE.
- DeVellis, R. F. (2016). *Scale development: theory and applications* (4th ed.). Los Angeles: SAGE.
- Dijkstra, J., van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2010). A new framework for designing programmes of assessment. *Advances in Health Sciences Education*, 15(3), 379–393. <https://doi.org/10.1007/s10459-009-9205-z>

- Doja, A., Horsley, T., & Sampson, M. (2014). Productivity in medical education research: An examination of countries of origin. *BMC Medical Education, 14*(1), 1–9. <https://doi.org/10.1186/s12909-014-0243-8>
- Dory, V., Gagnon, R., De Foy, T., Duyver, C., & Leconte, S. (2010). A novel assessment of an evidence-based practice course using an authentic assignment. *Med Teach, 32*(2), e65–70. <https://doi.org/10.3109/01421590903199718>
- Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830–837. <https://doi.org/10.1046/j.1365-2923.2003.01594.x>
- Driessen, E. (2017). Do portfolios have a future? *Advances in Health Sciences Education, 22*(1), 221–228. <https://doi.org/10.1007/s10459-016-9679-4>
- Driessen, E., & Scheele, F. (2013). What is wrong with assessment in postgraduate training? Lessons from clinical practice and educational research. *Medical Teacher, 35*(7), 569–574. <https://doi.org/10.3109/0142159X.2013.798403>
- Ebel, R. L. (1983). The Practical Validation of Tests of Ability. *Educational Measurement: Issues and Practice, 2*(2), 7–10. <https://doi.org/10.1111/j.1745-3992.1983.tb00688.x>
- Ellaway, R. H., Topps, D., & Pusic, M. (2019). Data, Big and Small: Emerging Challenges to Medical Education Scholarship, *94*(1), 31–36. <https://doi.org/10.1097/ACM.0000000000002465>
- Epstein, R. M. (2007). Medical education - Assessment in medical education. *New England Journal of Medicine, 356*(4), 387–396. <https://doi.org/10.1056/NEJMra054784>
- Eva, K. W., Bordage, G., Campbell, C., Galbraith, R., Ginsburg, S., Holmboe, E., & Regehr, G. (2015). Towards a program of assessment for health professionals: from training into practice. *Advances in Health Sciences Education, 21*(4), 1–17. <https://doi.org/10.1007/s10459-015-9653-6>
- Fitzgerald, J. T., Burkhardt, J. C., Kasten, S. J., Mullan, P. B., Santen, S. A., Sheets, K. J., ... Gruppen, L. D. (2016). Assessment challenges in competency-based education: A case study in health professions education. *Medical Teacher, 38*(5), 482–490. <https://doi.org/10.3109/0142159X.2015.1047754>
- Fortin, M.-F., & Gagnon, J. (2016). *Fondements et étapes du processus de recherche : méthodes quantitatives et qualitatives* (3rd ed.). Montréal, QC: Chenelière Éducation.
- Fotheringham, D. (2010). Triangulation for the assessment of clinical nursing skills: A review of theory, use and methodology. *International Journal of Nursing Studies, 47*(3), 386–391. <https://doi.org/10.1016/j.ijnurstu.2009.09.004>
- Frambach, J. M., van der vleuten, C. P., & Durning, S. J. (2013). AM Last Page: Quality Criteria in Qualitative and Quantitative Research. *Academic Medicine, 88*(4), 552. <https://doi.org/10.1097/ACM.0b013e31828abf7f>
- Frank, J. R., Sherbino, J., & Snell, L. (2018). Society Matters: Rethinking the Validity of Assessment in Med Ed. Key Literature in Medical Education PodCast. Retrieved April 14, 2019, from <https://keylimepodcast.libsyn.com/194-society-matters-rethinking-the-validity-of-assessment-in-med-ed>

- Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., Rossi, S. De, & Horsley, T. (2010). Toward a definition of competency-based education in medicine: a systematic review of published definitions. *Medical Teacher*, 32(8), 631–637.
- Frank, J. R., Snell, L., Englander, R., & Holmboe, E. S. (2017). Implementing competency-based medical education: Moving forward. *Medical Teacher*, 39(6). <https://doi.org/10.1080/0142159X.2017.1315069>
- Frenk, J., Chen, L., Bhutta, Z. A., Cohen, J., Crisp, N., Evans, T., ... Zurayk, H. (2010). Health professionals for a new century: transforming education to strengthen health systems in an interdependent world. *The Lancet*, 376(9756), 1923–1958. [https://doi.org/10.1016/S0140-6736\(10\)61854-5](https://doi.org/10.1016/S0140-6736(10)61854-5)
- Gallagher, F. (2014). La recherche descriptive interprétative. In M. Corbière & N. Larivière (Eds.), *Méthodes qualitatives, quantitatives et mixtes: dans la recherche en sciences humaines, sociales et de la santé* (pp. 5–27). Presses de l'Université du Québec.
- Gallagher, P., Pullon, S., Skinner, M., McHugh, P., McKinlay, E., & Gray, L. (2015). An interprofessional community education project as a socially accountable assessment. *Journal of Interprofessional Care*, 29(5), 509–511. Retrieved from <http://10.0.12.37/13561820.2015.1004040>
- Gibbs, G. (2019). How assessment frames student learning. In C. Bryan & K. Clegg (Eds.), *Innovative Assessment in Higher Education: A Handbook for Academic Practitioners* (2nd ed., pp. 22–35). Routledge.
- Gingerich, A., Kogan, J., Yeates, P., Govaerts, M., & Holmboe, E. (2014). Seeing the 'black box' differently: Assessor cognition from three research perspectives. *Medical Education*, 48(11), 1055–1068. <https://doi.org/10.1111/medu.12546>
- Gingerich, A., Ramlo, S. E., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2017). Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education*, 22(4), 819–838. <https://doi.org/10.1007/s10459-016-9711-8>
- Gingerich, A., van der Vleuten, C. P. M., Eva, K. W., & Regehr, G. (2014). More Consensus Than Idiosyncrasy. *Academic Medicine*, 89(11), 1510–1519. <https://doi.org/10.1097/ACM.0000000000000486>
- Ginsburg, S., McIlroy, J., Oulanova, O., Eva, K., & Regehr, G. (2010). Toward Authentic Clinical Evaluation: Pitfalls in the Pursuit of Competency. *Academic Medicine*, 85(5), 780–786. <https://doi.org/10.1097/ACM.0b013e3181d73fb6>
- Goodwin, L. D. (1997). Changing Conceptions of Measurement Validity. *Journal of Nursing Education*, 36(3), 102–107.
- Goodwin, L. D. (2002). Changing conceptions of measurement validity: an update on the new standards. *The Journal of Nursing Education*, 41(3), 100–106. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/11939227>
- Gotzmann, A., De Champlain, A., Touchie, C., & Boulet, J. (2013). *National Survey of the Physicians, Pharmacists, Nurses, and Public in Canada: 2013. Medical Council of Canada Blueprint Project*. Medical Council of Canada. Retrieved from <https://mcc.ca/media/Reports-2013-national-survey-1.pdf>

- Govaerts, M. (2015). Workplace-Based Assessment and Assessment for Learning: Threats to Validity. *Journal of Graduate Medical Education*, 7(2), 265–267. <https://doi.org/10.4300/JGME-D-15-00101.1>
- Govaerts, M. J. B., Schuwirth, L. W. T., van der Vleuten, C. P. M., & Muijtjens, A. M. (2011). Workplace-based assessment: Effects of rater expertise. *Advances in Health Sciences Education*, 16(2), 151–165. <https://doi.org/10.1007/s10459-010-9250-7>
- Govaerts, M., & van der Vleuten, C. P. M. (2013). Validity in work-based assessment: Expanding our horizons. *Medical Education*, 47(12), 1164–1174. <https://doi.org/10.1111/medu.12289>
- Grant, J. (2019). Principles of Curriculum Design. In T. Swanwick, K. Forrest, & B. C. O'Brien (Eds.), *Understanding Medical Education: Evidence, Theory and Practice* (3rd ed., pp. 71–88). Hoboken, NJ: John Wiley & Sons.
- Gruppen, L. D., Ten Cate, O., Lingard, L. A., Teunissen, P. W., & Kogan, J. R. (2018). Enhanced Requirements for Assessment in a Competency-Based, Time-Variable Medical Education System. *Academic Medicine: Journal Of The Association Of American Medical Colleges*, 93(3), S17–S21. <https://doi.org/10.1097/ACM.0000000000002066>
- Hanson, J. L., Rosenberg, A. A., & Lane, J. L. (2013). Narrative descriptions should replace grades and numerical ratings for clinical performance in medical education in the United States. *Frontiers in Psychology*, 4(668), 1–10. <https://doi.org/10.3389/fpsyg.2013.00668>
- Harden, R. M. (2015). Have we got assessment wrong? Thoughts from the Ottawa Conference and the need for a Pumpkin Plan. *Medical Teacher*, 37(3), 209–210. <https://doi.org/10.3109/0142159X.2015.1010497>
- Harris, P., Bhanji, F., Topps, M., Ross, S., Lieberman, S., Frank, J. R., ... Sherbino, J. (2017). Evolving concepts of assessment in a competency-based world. *Medical Teacher*, 39(6), 603–608. <https://doi.org/10.1080/0142159X.2017.1315071>
- Harrison, C., & Wass, V. (2016). The challenge of changing to an assessment for learning culture. *Medical Education*, 50(7), 704–706. <https://doi.org/10.1111/medu.13058>
- Hawkins, R. E., Welcher, C. M., Holmboe, E. S., Kirk, L. M., Norcini, J. J., Simons, K. B., & Skochelak, S. E. (2015). Implementation of competency-based medical education: Are we addressing the concerns and challenges? *Medical Education*, 49(11), 1086–1102. <https://doi.org/10.1111/medu.12831>
- Hays, R. (2008). Assessment in medical education: roles for clinical teachers. *The Clinical Teacher*, 5(1), 23–27. <https://doi.org/10.1111/j.1743-498X.2007.00165.x>
- Hodges, B. D. (2013). Assessment in the post-psychometric era: learning to love the subjective and collective. *Med Teach*, 35(7), 564–568. <https://doi.org/10.3109/0142159x.2013.789134>
- Hodges, B. D. (2015). Sea monsters & whirlpools: Navigating between examination and reflection in medical education. *Medical Teacher*, 37(3), 261–266. <https://doi.org/10.3109/0142159X.2014.993601>
- Hodges, B. D., Kuper, A., & Reeves, S. (2008). Discourse analysis. *BMJ (Clinical Research Ed.)*, 337(7669), a879. <https://doi.org/10.1136/bmj.a879>

- Hodges, B. D., & Lingard, L. (2012). *The question of competence: reconsidering medical education in the twenty-first century*. New York: ILR Press.
- Holmboe, E. S., Sherbino, J., Englander, R., Snell, L., & Frank, J. R. (2017). A call to action: The controversy of and rationale for competency-based medical education. *Medical Teacher, 39*(6). <https://doi.org/10.1080/0142159X.2017.1315067>
- Hong, Q. N., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., ... Pluye, P. (2018). The Mixed Methods Appraisal Tool (MMAT) version 2018 for information professionals and researchers. *Education for Information, 34*(4), 285–291. <https://doi.org/10.3233/EFI-180221>
- Huble, A. M., & Zumbo, B. D. (2011). Validity and the Consequences of Test Interpretation and Use. *Social Indicators Research, 103*(2), 219–230. <https://doi.org/10.1007/s11205-011-9843-4>
- Hunt, M. R. (2009). Strengths and challenges in the use of interpretive description: reflections arising from a study of the moral experience of health professionals in humanitarian work. *Qualitative Health Research, 19*(9), 1284–1292. <https://doi.org/10.1177/1049732309344612>
- Hupcey, J. E., Morse, J. M., Lenz, E. R., & Tason, M. C. (1996). Wilsonian methods of concept analysis: A critique. *Scholarly Inquiry for Nursing Practice: An International Journal, 10*(3), 185–210. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/9009818>
- Isaacson, J. J., & Stacy, A. S. (2009). Rubrics for clinical evaluation: objectifying the subjective experience. *Nurse Education in Practice, 9*(2), 134–140. <https://doi.org/10.1016/j.nepr.2008.10.015>
- Johnson, C. H. (2007). Competencies as an evaluation tool. *Clinics in Podiatric Medicine and Surgery, 24*(1), 103–117, vii. <https://doi.org/10.1016/j.cpm.2006.08.006>
- Kane, M. (2013). The Argument-Based Approach to Validation. *School Psychology Review, 42*(4).
- Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). Westport: American Council on Education/Praeger.
- Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73. <https://doi.org/10.1111/jedm.12001>
- Karpicke, J. D., & Grimaldi, P. J. (2012). Retrieval-Based Learning: A Perspective for Enhancing Meaningful Learning. *Educational Psychology Review, 24*(3), 401–418. <https://doi.org/10.1007/s10648-012-9202-2>
- Kasanda, C. D., Mitonga, K. H., Veii, K., & Zimba, R. F. (2013). Medical and pharmacy students' perceptions of the grading and assessment practices. *Frontiers in Psychology, 4*(423), 1–13. <https://doi.org/10.3389/fpsyg.2013.00423>
- Kealey, E. (2010). Assessment and Evaluation in Social Work Education: Formative and Summative Approaches. *Journal of Teaching in Social Work, 30*(1), 64–74. <https://doi.org/10.1080/08841230903479557>

- Kogan, J. R., Hatala, R., Hauer, K. E., & Holmboe, E. (2017). Guidelines: The do's, don'ts and don't knows of direct observation of clinical skills in medical education. *Perspectives On Medical Education*, 6(5), 286–305. <https://doi.org/10.1007/s40037-017-0376-7>
- Kogan, J. R., Holmboe, E. S., & Hauer, K. E. (2009). Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. *JAMA : The Journal of the American Medical Association*, 302(12), 1316–1326. <https://doi.org/10.1001/jama.2009.1365>
- Krueger, R. A., & Casey, M. A. (2015). *Focus groups: A practical Guide for Applied Research* (5th ed.). Thousand Oaks: Sage Publications.
- Labbé, M., Young, M., & Nguyen, L. H. P. P. (2018). Validity evidence as a key marker of quality of technical skill assessment in OTL-HNS. *The Laryngoscope*, 128, 2296–2300. <https://doi.org/10.1002/lary.27085>
- Lane, S. (2012). Consequences of Assessment and Accountability Systems Are Integral to the Argument-Based Approach to Validity. *Measurement: Interdisciplinary Research & Perspective*, 10(1–2), 71–74. <https://doi.org/10.1080/15366367.2012.679160>
- Larkins, S. L., Preston, R., Matte, M. C., Lindemann, I. C., Samson, R., Tandino, F. D., ... on behalf of the Training for Health (THEnet). (2013). Measuring social accountability in health professional education: Development and international pilot testing of an evaluation framework. *Medical Teacher*, 35, 32–45. <https://doi.org/10.3109/0142159X.2012.731106>
- Larsen, D. P., Butler, A. C., & Roediger, H. L. (2013). Comparative effects of test-enhanced learning and self-explanation on long-term retention. *Medical Education*, 47(7), 674–682. <https://doi.org/10.1111/medu.12141>
- Leavy, P. (2017). Mixed Methods Research Design. In P. Leavy (Ed.), *Research design : quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. https://doi.org/10.1007/978-3-642-35879-1_5
- LoBiondo-Wood, G., Haber, J., Cameron, C., & Singh, M. D. (2018). *Nursing Research in Canada: Methods, Critical Appraisal, and Utilization* (4th ed.). Toronto: Elsevier Canada.
- Lockyer, J., Carraccio, C., Chan, M.-K. K., Hart, D., Smee, S., Touchie, C., ... Frank, J. R. (2017). Core principles of assessment in competency-based medical education. *Medical Teacher*, 39(6), 609–616. <https://doi.org/10.1080/0142159X.2017.1315082>
- Lucey, C. R., Thibault, G. E., & Ten Cate, O. (2018). Competency-Based, Time-Variable Education in the Health Professions: Crossroads. *Academic Medicine: Journal Of The Association Of American Medical Colleges*, 93(3), S1–S5. <https://doi.org/10.1097/ACM.0000000000002080>
- MacLellan, E. (2004). Authenticity in assessment tasks: a heuristic exploration of academics' perceptions. *Higher Education Research & Development*, 23(1), 19–33. <https://doi.org/10.1080/0729436032000168478>
- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample Size in Qualitative Interview Studies. *Qualitative Health Research*, 26(13), 1753–1760. <https://doi.org/10.1177/1049732315617444>

- Marceau, M., Gallagher, F., Young, M., & St-Onge, C. (2018). Validity as a social imperative for assessment in health professions education: a concept analysis. *Medical Education*, 52(6), 641–653. <https://doi.org/10.1111/medu.13574>
- Marceau, M., St-Onge, C., Gallagher, F., & Young, M. (2019). Validity as a social imperative: Users' and leaders' perceived acceptability and anticipated feasibility. Submitted.
- Martinez, M. E., & Lipson, J. I. (1989). Assessment for Learning. *Educational Leadership*, 46, 73–75.
- Massie, J., & Ali, J. (2016). Workplace-based assessment: a review of user perceptions and strategies to address the identified shortcomings. *Advances in Health Sciences Education*, 21(2), 455–473.
- McAllister, S., Lincoln, M., Ferguson, A., & McAllister, L. (2013). Validating workplace performance assessments in health sciences students: a case study from speech pathology. *Journal of Applied Measurement*, 14(4), 356–374.
- McEwen, M., & Wills, E. M. (2019). *Theoretical Basis for Nursing* (5th ed.). Philadelphia: Wolters Kluwer.
- Mehrens, W. A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice*, 21(1), 16–18.
- Mertens, D. M. (2019). *Research and evaluation in education and psychology : integrating diversity with quantitative, qualitative, and mixed methods* (5th ed.). Thousand Oaks: Sage Publications.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York: Macmillan. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1002/j.2330-8516.1987.tb00244.x/abstract>
- Messick, S. (1995). Standards of Validity and the Validity of Standards in Performance Assessment. *Educational Measurement: Issues and Practice*, 14(4), 5–8. <https://doi.org/10.1111/j.1745-3992.1995.tb00881.x>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). Thousand Oaks: Sage Publications.
- Mislevy, R. J. (2007). Validity by Design. *Educational Researcher*, 36(8), 463–469. <https://doi.org/10.3102/0013189X07311660>
- Morgan, D. L. (2019). *Basic and advances focus groups*. Thousand Oaks: Sage Publications.
- Morse, J. M. (2000). Exploring Pragmatic Utility: Concept Analysis by Critically Appraising the Literature. In B. L. Rodgers & K. A. Knalf (Eds.), *Concept Development in Nursing* (2nd ed., pp. 333–352). Philadelphia: W. B. Saunders.
- Morse, J. M. (2015). Critical Analysis of Strategies for Determining Rigor in Qualitative Inquiry. *Qualitative Health Research*, 25(9), 1212–1222. <https://doi.org/10.1177/1049732315588501>
- Morse, J. M., Niehaus, L., Wolfe, R. R., & Wilkins, S. (2006). *The role of the theoretical drive in maintaining validity in mixed method research*. (August), 6–10.

- Murphy, D. J., Bruce, D., & Eva, K. W. (2008). Workplace-based assessment for general practitioners: Using stakeholder perception to aid blueprinting of an assessment battery. *Medical Education*, 42(1), 96–103. <https://doi.org/10.1111/j.1365-2923.2007.02952.x>
- Naumanen-Tuomela, P. (2001). Concept analysis of expertise of occupational health nurses applying Rodgers's evolutionary model. *International Journal of Nursing Practice*, 7(4), 257–265. Retrieved from <http://www.scopus.com/inward/record.url?eid=2-s2.0-0035436604&partnerID=40&md5=e6b7882b18c0f2d8f11310c542177d58>
- Neary, M. (2001). Responsive assessment: assessing student nurses' clinical competence. *Nurse Education Today*, 21(1), 3–17. <https://doi.org/10.1054/nedt.2000.0508>
- Norcini, J.J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206–214. <https://doi.org/10.3109/0142159X.2011.551559>
- Norcini, J.J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... Roberts, T. (2011). Criteria for good assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 206–214. <https://doi.org/10.3109/0142159X.2011.551559>
- Norcini, J. J., & McKinley, D. W. (2007). Assessment methods in medical education. *Teaching & Teacher Education*, 23(3), 239–250. <https://doi.org/10.1016/j.tate.2006.12.021>
- Norcini, J. J., & Zaidi, Z. (2019). Workplace Assessment. In T. Swanwick, K. Forrest, & B. C. O'Brien (Eds.), *Understanding Medical Education: Evidence, Theory and Practice* (3rd ed., pp. 319–334). Hoboken, NJ: John Wiley & Sons.
- Norman, G., Norcini, J., & Bordage, G. (2014). Competency-Based Education: Milestones or Millstones 1 ? *Journal of Graduate Medical Education*, 6(1), 1–6. <https://doi.org/10.4300/jgme-d-13-00445.1>
- O'Cathain, A. (2010). Assessing the quality of mixed methods research. In Abbas Tashakkori & C. Teddlie (Eds.), *Sage Handbook of Mixed Methods Research* (2nd ed., pp. 531–555). Los Angeles: Sage Publications.
- Onyura, B., Légaré, F., Baker, L., Reeves, S., Rosenfield, J., Kitto, S., ... Leslie, K. (2015). Affordances of Knowledge Translation in Medical Education. *Academic Medicine*, 90(4), 518–524. <https://doi.org/10.1097/ACM.0000000000000590>
- Palm, T. (2008). Performance Assessment and Authentic Assessment: A conceptual Analysis of the Literature. *Practical Assessment Research Evaluation*, 13(4), 1–24.
- Patton, M. Q. (2015). *Qualitative Research & Evaluation Methods* (4th ed.). Thousand Oaks: Sage Publications.
- Peeters, M. J. (2017). Targeting assessment for learning within pharmacy education. *American Journal of Pharmaceutical Education*, 81(8), 5–9. <https://doi.org/10.5688/ajpe6243>

- Pennaforte, T., & Loye, N. (2017). Une approche pragmatique de validation en éducation médicale : L'application du modèle de Kane à un outil d'évaluation du raisonnement clinique. In E. Dionne & I. Raïche (Eds.), *Mesure et évaluation des compétences en éducation médicale* (pp. 143–176). Presses de l'Université du Québec.
- Peters, H., Holzhausen, Y., Boscardin, C., ten Cate, O., & Chen, H. C. (2017). Twelve tips for the implementation of EPAs for assessment and entrustment decisions. *Medical Teacher*, 39(8), 802–807. <https://doi.org/10.1080/0142159X.2017.1331031>
- Popham, W. (1997). Consequential validity: Right concern-wrong concept. *Educational Measurement: Issues and Practice*, 16(2), 9–13. <https://doi.org/10.1111/j.1745-3992.1997.tb00586.x>
- Prégent, R., Bernard, H., & Kozanitis, A. (2009). *Enseigner à l'université dans une approche-programme*. Montréal: Presses internationales Polytechnique.
- Prideaux, D., Roberts, C., Eva, K., Centeno, A., Mccrorie, P., Mcmanus, C., ... Wilkinson, D. (2011). Assessment for selection for the health care professions and specialty training: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 215–223. <https://doi.org/10.3109/0142159X.2011.551560>
- Pugh, D., & Regehr, G. (2016). Taking the sting out of assessment: is there a role for progress testing? *Medical Education*, 50(7), 721–729. <https://doi.org/10.1111/medu.12985>
- Pugh, D., Touchie, C., Wood, T. J., & Humphrey-Murto, S. (2014). Progress testing: Is there a role for the OSCE? *Medical Education*, 48(6), 623–631. <https://doi.org/10.1111/medu.12423>
- Rentschler, D., Eaton, J., Cappiello, J., McNally, S., & McWilliam, P. (2007). Evaluation of undergraduate students using objective structured clinical evaluation. *Journal of Nursing Education*, 46(3), 135–139.
- Ricci, M., St-Onge, C., Xiao, J., & Young, M. (2018). Students as stakeholders in assessment: how students perceive the value of an assessment. *Perspectives on Medical Education*, 7(6), 352–361. <https://doi.org/10.1007/s40037-018-0480-3>
- Richardson, L., Resick, L., Leonardo, M., & Pearsall, C. (2009). Undergraduate students as standardized patients to assess advanced practice nursing student competencies. *Nurse Educator*, 34(1), 12–16. <https://doi.org/http://dx.doi.org/10.1097/01.NNE.0000343397.26695.89>
- Roberts, C., Newble, D., Jolly, B., Reed, M., & Hampton, K. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher*, 28(6), 535–543. <https://doi.org/10.1080/01421590600711187>
- Rodgers, B. L. (1989). Concepts, analysis and the development of nursing knowledge: the evolutionary cycle. *Journal of Advanced Nursing*, 14(4), 330–335. <https://doi.org/10.1111/j.1365-2648.1989.tb03420.x>
- Rodgers, B. L., & Knafl, K. A. (2000). *Concept development in nursing* (2nd ed.). Philadelphia: W. B. Saunders.

- Rosen, M. a, Salas, E., Silvestri, S., Wu, T. S., & Lazzara, E. H. (2008). A measurement tool for simulation-based training in emergency medicine: the simulation module for assessment of resident targeted event responses (SMARTER) approach. *Simulation in Healthcare : Journal of the Society for Simulation in Healthcare*, 3(3), 170–179. <https://doi.org/10.1097/SIH.0b013e31817fe513>
- Royal College of Physicians and Surgeons of Canada. (2014). *Competence by Design (CBD)*. Royal College of Physicians and Surgeons of Canada. Retrieved from http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/cbd/what_is_cbd_e.pdf
- Royal College of Physicians and Surgeons of Canada. (2015). *CanMEDS 2015 Physician Competency Framework*. (J.R. Frank, L. Snell, & J. Sherbino, Eds.). Ottawa: Royal College of Physicians and Surgeons of Canada.
- Royal College of Physicians and Surgeons of Canada. (2019a). CBD and assessment. Retrieved March 20, 2019, from <http://www.royalcollege.ca/rcsite/cbd/cbd-assessment-e>
- Royal College of Physicians and Surgeons of Canada. (2019b). EPAs and Milestones. Retrieved March 20, 2019, from <http://www.royalcollege.ca/rcsite/cbd/implementation/cbd-milestones-epas-e>
- Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. St-Laurent: Éditions du Renouveau pédagogique Inc.
- Schultz, K., Griffiths, J., & Lacasse, M. (2015). The Application of Entrustable Professional Activities to Inform Competency Decisions in a Family Medicine Residency Program. *Academic Medicine: Journal Of The Association Of American Medical Colleges*, 90(7), 888–897. <https://doi.org/10.1097/ACM.0000000000000671>
- Schut, S., Driessen, E., van Tartwijk, J., van der Vleuten, C., & Heeneman, S. (2018). Stakes in the eye of the beholder: an international study of learners' perceptions within programmatic assessment. *Medical Education*, 52(6), 654–663. <https://doi.org/10.1111/medu.13532>
- Schuwirth, L., & Ash, J. (2013). Assessing tomorrow's learners: in competency-based education only a radically different holistic method of assessment will work. Six things we could forget. *Medical Teacher*, 35(7), 555–559. <https://doi.org/10.3109/0142159X.2013.787140>
- Schuwirth, L. W. T., & van der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, 33(1), 478–485. <https://doi.org/10.3109/0142159X.2011.565828>
- Schuwirth, L.W.T., & van der Vleuten, C. P. M. (2006). A plea for new psychometric models in educational assessment. *Medical Education*, 40(4), 296–300. <https://doi.org/10.1111/j.1365-2929.2006.02405.x>
- Schuwirth, L.W.T., Colliver, J., Gruppen, L., Kreiter, C., Mennin, S., Onishi, H., ... Wagner-Menghin, M. (2011). Research in assessment: Consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, 33(3), 224–233. Retrieved from <http://10.0.12.37/0142159X.2011.551558>

- Schuwirth, L.W.T., & van der Vleuten, C. P. M. (2012). Programmatic assessment and Kane's validity perspective. *Medical Education*, 46(1), 38–48. <https://doi.org/10.1111/j.1365-2923.2011.04098.x>
- Schuwirth, L.W.T., & van der Vleuten, C. P. M. (2019). How to Design a Useful Test: The Principles of Assessment. In T. Swanwick, K. Forrest, & B. C. O'Brien (Eds.), *Understanding Medical Education: Evidence, Theory and Practice* (3rd ed., pp. 277–289). Hoboken, NJ: Wiley Blackwell.
- Schuwirth, L.W.T., & van der Vleuten, C. P. M. (2013). Assessment. In T. Swanwick (Ed.), *Understanding Medical Education: Evidence, Theory and Practice*. John Wiley & Sons, Incorporated. Retrieved from <http://search.jamas.or.jp/link/ui/2013104810>
- Scoles, P. V, Hawkins, R. E., & LaDuca, A. (2003). Assessment of Clinical Skills in Medical Practice. *Journal of Continuing Education in the Health Professions*, 23(3), 182–190.
- Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*, 16(2), p5-8,13,24.
- Shumway, J. M., & Harden, R. M. (2003). AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Medical Teacher*, 25(6), 569–584. <https://doi.org/10.1080/0142159032000151907>
- Sidani, S., & Braden, C. J. (2011). *Design, Evaluation, and Translation of Nursing Interventions*. Chichester: Wiley-Blackwell. <https://doi.org/10.1002/9781118785553.ch12>
- Sireci, S. G., & Sukin, T. (2013). Test validity. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbooks in psychology. APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (pp. 61–84). Washington: American Psychological Association. <https://doi.org/10.1037/14047-004>
- Sklar, D. P. (2015). Competencies, Milestones, and Entrustable Professional Activities. *Academic Medicine*, 90(4), 395–397. <https://doi.org/10.1097/acm.0000000000000659>
- St-Onge, C., Chamberland, M., Lévesque, A., & Varpio, L. (2016). Expectations, observations, and the cognitive processes that bind them: expert assessment of examinee performance. *Advances in Health Sciences Education*, 21(3), 627–642. <https://doi.org/10.1007/s10459-015-9656-3>
- St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education*, 22(4), 853–867. <https://doi.org/doi:10.1007/s10459-016-9716-3>
- Stevens, B., Hyde, J., Knight, R., Shires, A., & Alexander, R. (2017). Competency-based training and assessment in Australian postgraduate clinical psychology education. *Clinical Psychologist*, 21(3), 174–185. Retrieved from <http://10.0.4.87/cp.12061>
- Streiner, D. L., Norman, G. R., & Cairney, J. (2015). *Health measurement scales* (5th ed.). Oxford: Oxford University Press.

- Surjadi, M., Stringari-Murray, S., & Saxe, J. A. M. (2019). Entrustable Professional Activities in Nurse Practitioner Education. *Journal for Nurse Practitioners*, 15(5), e97–e102. <https://doi.org/10.1016/j.nurpra.2018.12.030>
- Tardif, J. (2006). *L'évaluation des compétences: Documenter le parcours de développement*. Montréal: Chenelière éducation.
- Tashakkori, A., & Teddlie, C. (2008). Quality of inferences in mixed methods research: Calling for an integrative framework. In M. Bergman (Ed.), *Advances in mixed methods research: Theories and applications*. London: Sage.
- Taylor, C. S. (2013). *Validity and validation*. New York: Oxford University Press.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research : integrating quantitative and qualitative approaches in the social and behavioral sciences*. Thousand Oaks: Sage Publications.
- Tekian, A., Hodges, B. D., Roberts, T. E., Schuwirth, L., & Norcini, J. (2015). Assessing competencies using milestones along the way. *Medical Teacher*, 37(4), 399–402. <https://doi.org/10.3109/0142159X.2014.993954>
- Tekian, A., Watling, C. J., Roberts, T. E., Steinert, Y., & Norcini, J. (2017). Qualitative and quantitative feedback in the context of competency-based education. *Medical Teacher*, 39(12). <https://doi.org/10.1080/0142159X.2017.1372564>
- Ten Cate, O. (2013). Nuts and Bolts of Entrustable Professional Activities. *Journal of Graduate Medical Education*, 5(1), 157–158. <https://doi.org/10.1007/s11606-012-2193-3>
- Thorne, S. (2016). *Interpretative description: qualitative research for applied practice*. New York: Routledge.
- Thorne, S., Reimer Kirkham, S., & O'Flynn-Magee, K. (2004). The Analytic Challenge in Interpretive Description. *International Journal of Qualitative Methods*, 3(1), 1–11. <https://doi.org/10.1287/trsc.1090.0266>
- Timmerman, A. A., & Dijkstra, J. (2017). A practical approach to programmatic assessment design. *Advances in Health Sciences Education*, 22(5), 1169–1182. <https://doi.org/10.1007/s10459-017-9756-3>
- Toftagen, R., & Fagerstrøm, L. M. (2010). Rodgers' evolutionary concept analysis - a valid method for developing knowledge in nursing science. *Scandinavian Journal of Caring Sciences*, 24, 21–31. <https://doi.org/10.1111/j.1471-6712.2010.00845.x>
- Touchie, C. (2013). Report of the Incidence and Prevalence of Diseases and Other Health Related Issues in Canada: A study for the MCC Blueprint Project. Retrieved August 14, 2019, from <https://mcc.ca/wp-content/uploads/Touchie-Incidence-Prevalence-Report.pdf>
- Tracy, S. J. (2013). *Qualitative research methods*. Chichester: Wiley-Blackwell.
- Ulfvarson, J., & Oxelmark, L. (2012). Developing an assessment tool for intended learning outcomes in clinical practice for nursing students. *Nurse Education Today*, 32(6), 703–708. <https://doi.org/10.1016/j.nedt.2011.09.010>
- van der Vleuten, C. P. M., & Dannefer, E. F. (2012). Towards a systems approach to assessment. *Medical Teacher*, 34(3), 185–186. <https://doi.org/10.3109/0142159X.2012.652240>

- van der Vleuten, C. P. M., & Heeneman, S. (2016a). A new holistic way of assessment: programmatic assessment. *Revista de La Fundación Educación Médica*, 19(6), 275–279. <https://doi.org/10.33588/fem.196.860>
- van der Vleuten, C. P. M., & Heeneman, S. (2016b). On the issue of costs in programmatic assessment. *Perspectives on Medical Education*, 5(5), 303–307. Retrieved from <http://10.0.3.239/s40037-016-0295-z>
- van der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: from methods to programmes. *Medical Education*, 39(3), 309–317. <https://doi.org/10.1111/j.1365-2929.2005.02094.x>
- van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Dijkstra, J., Tigelaar, D., Baartman, L. K. J., & van Tartwijk, J. (2012). A model for programmatic assessment fit for purpose. *Medical Teacher*, 34(3), 205–214. <https://doi.org/10.3109/0142159X.2012.652239>
- van der Vleuten, C. P. M., Schuwirth, L. W. T., Driessen, E. W., Govaerts, M. J. B., & Heeneman, S. (2015). Twelve Tips for programmatic assessment. *Medical Teacher*, 37(7), 641–646. <https://doi.org/10.3109/0142159X.2014.973388>
- van der Vleuten, C. P. M., Schuwirth, L. W. T., Scheele, F., Driessen, E. W., & Hodges, B. (2010). The assessment of professional competence: building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*, 24(6), 703–719. <https://doi.org/10.1016/j.bpobgyn.2010.04.001>
- Van Tartwijk, J., & Driessen, E. W. (2009). Portfolios for assessment and learning: AMEE Guide no. 45. *Medical Teacher*, 31(9), 790–801. <https://doi.org/10.1080/01421590903139201>
- Varpio, L., Ajjawi, R., Monrouxe, L. V., O'Brien, B. C., & Rees, C. E. (2017). Shedding the cobra effect: Problematising thematic emergence, triangulation, saturation and member checking. *Medical Education*, 51(1), 40–50. <https://doi.org/10.1111/medu.13124>
- Varpio, L., Paradis, E., Uijtdehaage, S., & Young, M. (n.d.). The distinctions between Theory, Theoretical Framework, and Conceptual Framework. *Academic Medicine*.
- Vriend Van Duinen, D. (2005). Authentic Assessment. *International Journal of Learning*, 12(6), 141–148. Retrieved from <http://search.ebscohost.com/login.aspx?direct=true&db=e0h&AN=27577900&site=ehost-live>
- Wagner, S. J., & Reeves, S. (2015). Milestones and entrustable professional activities: The key to practically translating competencies for interprofessional education? *Journal of Interprofessional Care*, 29(5), 507–508. <https://doi.org/10.3109/13561820.2014.1003636>
- Walker, L. O., & Avant, K. C. (2019). *Strategies for Theory Construction in Nursing* (6th ed.). New York: Pearson.
- Wass, V., Van der Vleuten, C., Shatzer, J., & Jones, R. (2001). Assessment of clinical competence. *Lancet*, 357(9260), 945–949. [https://doi.org/10.1016/S0140-6736\(00\)04221-5](https://doi.org/10.1016/S0140-6736(00)04221-5)
- Watling, C. J., & Ginsburg, S. (2019). Assessment, feedback and the alchemy of learning. *Medical Education*, 53(1). <https://doi.org/10.1111/medu.13645>

- Wen, L. S., Greysen, S. R., Keszthelyi, D., Bracero, J., & de Roos, P. D. G. (2011). Social accountability in health professionals' training. *The Lancet*, 378(9807), e12–e13. [https://doi.org/10.1016/S0140-6736\(10\)62314-8](https://doi.org/10.1016/S0140-6736(10)62314-8)
- Wiggins, G. (1989). A True Test: Toward More Authentic and Equitable Assessment. *Phi Delta Kappan*, 70(May), 703–713. <https://doi.org/10.1177/003172171109200721>
- Wiggins, G. (2011). A true test. *Kappan*, 92(7), 81–93. Retrieved from http://www.nisdx.org/cms/lib/TX21000351/Centricity/Domain/21/jcarlisle/ATrueTest_Wiggins.pdf
- Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37(1), 3–14. <https://doi.org/10.1016/j.stueduc.2011.03.001>
- Wills, E., & McEwen, M. (2011). Concept Development: Clarifying Meaning of Terms. In M. McEwen & E. Wills (Eds.), *Theoretical Basis for Nursing* (3rd ed., pp. 46–67). Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins Health.
- Wilson, J. (1963). *Thinking with Concepts*. Cambridge: Cambridge University Press.
- Woo, K. (2019). *Polit & Beck : Canadian essentials of nursing research* (4th ed.). Philadelphia: Wolters Kluwer.
- Wood, D. F. (2019). Formative Assessment: Assessment for Learning. In T. Swanwick, K. Forrest, & B. C. O'Brien (Eds.), *Understanding Medical Education: Evidence, Theory and Practice* (3rd ed., pp. 361–373). Hoboken, NJ: John Wiley & Sons.
- Wrigley, W., van der Vleuten, C. P. M., Freeman, A., & Muijtjens, A. (2012). A systemic framework for the progress test: Strengths, constraints and issues: AMEE Guide No. 71. *Medical Teacher*, 34(9), 683–697. <https://doi.org/10.3109/0142159X.2012.704437>
- Young, M., St-Onge, C., Xiao, J., Vachon Lachiver, E., & Torabi, N. (2018). Characterizing the literature on validity and assessment in medical education: a bibliometric study. *Perspectives on Medical Education*, 7(3), 182–191. <https://doi.org/10.1007/s40037-018-0433-x>

ANNEXES

Annexe 1. Approbation éthique de l'Université de Sherbrooke



Comité d'éthique de la recherche
Éducation et sciences sociales

Attestation de conformité

Le comité d'éthique de la recherche Éducation et sciences sociales de l'Université de Sherbrooke certifie avoir examiné la proposition de recherche suivante :

La validité en tant qu'impératif social dans un contexte d'évaluation des apprentissages

Mélanie Marceau

Étudiante, Doctorat en sciences cliniques, Faculté de médecine et des sciences de la santé

Le comité estime que la recherche proposée est conforme aux principes éthiques énoncés dans la *Politique en matière d'éthique de la recherche avec des êtres humains (2500-028)*.

Membres du comité

Eric Yergeau, président du comité, professeur à la Faculté d'éducation, Département d'orientation professionnelle

Suzanne Guillemette, professeure à la Faculté d'éducation, Département de gestion de l'éducation et de la formation

Sawsen Lakhai, professeure à la Faculté d'éducation, Département de pédagogie

Mélanie Lapalme, professeure à la Faculté d'éducation, Département de psychoéducation

Yves Lenoir, professeur à la Faculté d'éducation, Département d'enseignement au préscolaire et au primaire

Julie Myre-Bisailon, professeure à la Faculté d'éducation, Département d'adaptation scolaire et sociale

Carlo Spallanzani, professeur à la Faculté d'éducation physique et sportive

Marianne Xhignesse, professeure à la Faculté de médecine et des sciences de la santé, Département de médecine de famille

Vincent Beaucher, membre versé en éthique

France Dupuis, membre de la collectivité

Le présent certificat est valide pour la durée de la recherche, à condition que la personne responsable du projet fournisse au comité un rapport de suivi annuel, faute de quoi le certificat peut être révoqué.

Le président du comité,

Eric Yergeau, 22 juin 2016

Annexe 2. Approbation éthique de l'Université McGill

Bonjour Madame Coulombe,

Le projet CER-ESS-2016-34 de la candidate au doctorat en sciences, Mélanie Marceau, s'inscrit dans l'Entente CRÉPUQ et, à cette fin, nous reconnaissons l'évaluation effectuée par le CÉR de Université de Sherbrooke.

Dr. Young : Please keep this email as confirmation for your files.

Merci et bonne journée,
Ilde

Ilde Lepore

Ethics Officer / Agente d'éthique

Faculty of Medicine Research, Graduate Studies and IRB /Recherche, Études supérieures et CÉR

Faculté de médecine|

Université McGill University

Bonjour,

Une de nos étudiantes au doctorat en sciences cliniques (pédagogie des sciences de la santé) réalise un projet de recherche qui prévoit le recrutement de professeurs, chargés de cours, chargés d'enseignement, directeur d'un programme ou encore membres du comité du curriculum d'un programme professionnalisant en santé dans quatre universités québécoises (Université de Montréal, Université Laval, Université McGill et Université de Sherbrooke).

Comme pour le projet d'un étudiant au doctorat en éducation que je vous avais transmis en février dernier (Université Laval et Université de Montréal), il ne s'agit pas d'une recherche admissible à l'Entente CREPUQ puisqu'il n'y a pas de co-chercheurs dans vos universités. Toutefois, à l'Université McGill, l'une des directrices de l'étudiante provient de votre établissement (D^r Meredith Young).

Comme convenu pour le projet transmis en février 2016, je vous fais parvenir le dossier de Mme Mélanie Marceau, étudiante au doctorat en sciences cliniques. Je joins le projet dans sa version finale, la réponse du comité d'éthique et l'approbation finale du comité.

Je vous serais reconnaissante de guider notre étudiante dans les démarches supplémentaires qui pourraient être nécessaires pour l'analyse de son dossier par votre comité d'éthique de la recherche.

Je vous remercie à l'avance pour votre précieuse aide et je vous souhaite une excellente journée

Carole

Carole Coulombe

Coordonnatrice à l'éthique de la recherche

Comité d'éthique de la recherche – Lettres et sciences humaines

Comité d'éthique de la recherche – Éducation et sciences sociales

Service d'appui à la recherche, à l'innovation et à la création

Université de Sherbrooke

Annexe 3. Approbation éthique de l'Université Laval

	
<small>Vice-rectorat à la recherche et à la création Comité d'éthique de la recherche</small>	
<h2>APPROBATION DE L'ÉTHIQUE</h2>	
<p>Projet de recherche impliquant des êtres humains ou la consultation de renseignements personnels</p>	
<p>Ce projet de recherche a été examiné en conformité avec les <i>Modalités de gestion de l'éthique de la recherche sur des êtres humains</i> de l'Université Laval, par le Comité sectoriel d'éthique de la recherche en sciences de la santé</p>	
Projet intitulé :	La validité en tant qu'impératif social dans un contexte d'évaluation des apprentissages
Nom du chercheur :	Madame Mélanie Marceau
Nom du directeur de recherche :	Madame Frances Gallagher
Numéro d'approbation :	2016-220 / 30-08-2016
Date de décision :	30 août 2016
Date d'expiration de l'approbation :	1 ^{er} septembre 2017
<p>Après examen des informations et des documents qui lui ont été transmis, le Comité a constaté que ce projet respecte les principes d'éthique de la recherche avec des êtres humains. Il prend acte de la confirmation écrite de la chercheuse à l'effet qu'elle a pris connaissance des mesures de suivi¹ associées à l'émission de l'approbation éthique de son projet et qu'elle accepte de les appliquer. Par conséquent, le Comité approuve ce projet pour un an.</p>	
<hr/> Mahmoud Rouabhia , président Comité d'éthique de la recherche en sciences de la santé	<u>8 septembre 2016</u> Date
<hr/> <small>¹ Rappel des mesures de suivi au verso</small>	

Annexe 4. Approbation éthique de l'Université de Montréal



Comité plurifacultaire d'éthique de la recherche

7 septembre 2016

Madame Mélanie Marceau
 Candidate au doctorat
 Sciences cliniques, pédagogie de la santé, Université de Sherbrooke

OBJET: Approbation éthique

Mme Mélanie Marceau,

Le *Comité plurifacultaire d'éthique de la recherche (CPER)* a étudié le projet de recherche intitulé « La validité en tant qu'impératif social dans un contexte d'évaluation des apprentissages » et a délivré le certificat d'éthique demandé suite à la satisfaction des exigences précédemment émises.

Notez qu'il y apparaît une mention relative à un suivi annuel et que le certificat comporte une date de fin de validité. En effet, afin de répondre aux exigences éthiques en vigueur au Canada et à l'Université de Montréal, nous devons exercer un suivi annuel auprès des chercheurs et étudiants-chercheurs.

De manière à rendre ce processus le plus simple possible et afin d'en tirer pour tous le plus grand profit, nous avons élaboré un court questionnaire qui vous permettra à la fois de satisfaire aux exigences du suivi et de nous faire part de vos commentaires et de vos besoins en matière d'éthique en cours de recherche. Ce questionnaire de suivi devra être rempli annuellement jusqu'à la fin du projet et pourra nous être retourné par courriel. La validité de l'approbation éthique est conditionnelle à ce suivi. Sur réception du dernier rapport de suivi en fin de projet, votre dossier sera clos.

Il est entendu que cela ne modifie en rien l'obligation pour le chercheur, tel qu'indiqué sur le certificat d'éthique, de signaler au CPER tout incident grave dès qu'il survient ou de lui faire part de tout changement anticipé au protocole de recherche.

Nous vous prions d'agréer, Madame, l'expression de nos sentiments les meilleurs,

Jeán Pouparç, Président
Comité plurifacultaire d'éthique de la recherche (CPER)
 Université de Montréal

JP/RS/rs

c.c. Gestion des certificats, BRDV
 Frances Gallagher, professeure, École de sciences infirmières
 Christina St-Onge, professeure, Département de médecine
 p.j. Certificat CPER-16-084-D

Annexe 5. Invitation à participer
Personnel enseignant et membres de comités d'évaluation des apprentissages

INVITATION À PARTICIPER AU PROJET DE RECHERCHE :

**La validité en tant qu'impératif social
dans le contexte de l'évaluation des apprentissages**

Mélanie Marceau, inf. M.Sc.
Candidate au doctorat en sciences cliniques (pédagogie des sciences de la santé)
Université de Sherbrooke

Équipe de direction :
Frances Gallagher¹, Ph.D., Christina St-Onge¹, Ph.D., Meredith Young², Ph.D.
¹Faculté de médecine et sciences de la santé, Université de Sherbrooke
²Faculté de médecine, Université McGill

Madame,
Monsieur,

Vous êtes invité(e) à participer à la recherche en titre. Le but de cette étude sera d'explorer le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages auprès de programmes universitaires en santé au Québec (ergothérapie, physiothérapie, médecine et sciences infirmières).

Les participants recherchés sont les professeurs et professeures, les chargés et chargées de cours ainsi que les chargés et chargées d'enseignement qui jouent un rôle dans la conception, l'élaboration ou le monitoring de l'évaluation des apprentissages. Ce projet a été approuvé par les comités d'éthique des universités visées par la recherche (Sherbrooke, Laval, McGill et Montréal)

Pourquoi s'intéresser à la validité de l'évaluation des apprentissages?

La mesure de la qualité (validité) des stratégies d'évaluation des apprentissages est essentielle pour assurer une interprétation juste et appropriée des résultats de l'évaluation, car ces résultats peuvent avoir des conséquences importantes, tant pour les apprenants que pour la société. Dans cette étude, nous nous intéressons à la qualité de l'évaluation des apprentissages dans une approche par compétences (APC) qui semble privilégiée dans les programmes universitaires en sciences de la santé.

Mais qu'est-ce que la validité en tant qu'impératif social?

Une récente analyse de discours effectuée par des membres de notre équipe de recherche a mis en évidence des insatisfactions concernant les stratégies de validation disponibles (St-Onge et al., 2017) ainsi qu'une nouvelle conceptualisation de la validité, soit **la validité en tant qu'impératif social**.

Comme notre compréhension actuelle de la validité en tant qu'impératif social est fondée principalement sur l'analyse de documents, ce concept émergent demande une exploration beaucoup plus approfondie, notamment, pour vérifier son acceptabilité auprès des personnes concernées directement par l'évaluation des apprentissages.

En quoi consiste votre participation au projet?

Votre contribution à ce projet de recherche consiste à participer à une entrevue individuelle qui se déroulera en présentiel à la l'Université Laval (lieu à déterminer). Le temps consacré à cette recherche, soit environ 90 minutes, est le principal inconvénient lié à votre participation.

Lors de l'entrevue, des explications sur le contexte du projet de recherche (analyse de discours et analyse de concept) seront présentées au tout début de la rencontre. Puis, vous serez invité à compléter le formulaire de consentement et un questionnaire sociodémographique.

Qu'est-ce que la chercheuse fera avec les données recueillies?

Pour éviter votre identification comme participant dans ce projet, les données seront traitées de manière entièrement confidentielle. Votre nom sera remplacé par un code numérique qui sera connu seulement de l'équipe de recherche et qui sera maintenu sous clé. Le système de codage sera détruit à la fin de l'étude. Les jugements et opinions partagés sur les questionnaires demeureront strictement anonymes. Cinq ans après la publication des résultats, les documents papier et électroniques seront détruits. Les données ne seront pas utilisées à d'autres fins que celles décrites dans le présent document : thèse, articles ou présentations scientifiques. Comme il s'agit d'une étude de nature exploratoire, l'équipe de recherche pourrait avoir besoin de retourner aux données recueillies pour approfondir l'analyse. À tout moment, les données demeureront anonymes et confidentielles.

Est-il obligatoire de participer?

Non. La participation à cette étude se fait sur une base volontaire. Vous êtes **entièrement libre de participer ou non**, et de vous retirer en tout temps sans avoir à motiver votre décision ni à subir de préjudice de quelque nature que ce soit.

Y a-t-il des risques, inconvénients ou bénéfiques?

Au-delà des inconvénients mentionnés jusqu'ici (temps consacré à la recherche), les chercheuses considèrent que les risques possibles sont minimaux. La contribution à l'avancement des connaissances au sujet de la validité en tant qu'impératif social sont les bénéfices prévus. Aucune compensation d'ordre monétaire n'est offerte pour la participation à ce projet.

Que faire si j'ai des questions concernant le projet?

Si vous avez des questions concernant ce projet de recherche, n'hésitez pas à communiquer avec moi aux coordonnées indiquées ci-dessous.

Mélanie Marceau, inf. M.Sc.

Candidate au doctorat en sciences cliniques (pédagogie des sciences de la santé)

Faculté de médecine et des sciences de la santé

Université de Sherbrooke

[XXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

Annexe 6. Questionnaire sociodémographique
Personnel enseignant et membres de comités d'évaluation des apprentissages

1. Êtes-vous :

Un homme?

Une femme?

2. Quel âge avez-vous? _____ ans

3. Quelle est votre discipline professionnelle?

Sciences infirmières

Médecine

Physiothérapie

Ergothérapie

Éducation

Autre : _____

4. Quel emploi occupez-vous au sein de votre université?

Professeur ou professeure

Professeur ou professeure de clinique

Chargé ou chargée d'enseignement

Chargé ou chargée de cours

Coordonnateur ou coordonnatrice de stage

Conseiller ou conseillère pédagogique

Autre : _____

5. Depuis combien de temps occupez-vous cette fonction? _____ années

6. Êtes-vous impliqué davantage :

Au 1^{er} cycle universitaire?

Au 2^e cycle universitaire?

Au 1^{er} et au 2^e cycles universitaires?

7. Quelle est votre contribution au sujet de l'évaluation des apprentissages dans votre programme? (cocher toutes les réponses qui s'appliquent)

- Je développe mes propres évaluations
- Je collabore avec des collègues au développement des stratégies d'évaluation
- Je participe à la validation des stratégies d'évaluation
- Je suis directeur ou directrice de programme
- Je suis membre d'un comité de programme
- Je suis membre d'un comité d'évaluation
- Autres : _____

8. Depuis combien d'années êtes-vous impliqué dans l'évaluation des apprentissages auprès des étudiants? _____ années

9. Quelles sont les principales stratégies d'évaluation des apprentissages utilisées dans votre cours et/ou programme? (cocher toutes les réponses qui s'appliquent)

- Questionnaire à choix multiples (QCM)
- Questions à réponse courte (QRC)
- Questions à développement
- Travail écrit
- Examen clinique objectif structuré (ECOS)
- Examen oral structuré
- Rubrique d'évaluation
- Portfolio
- Carte conceptuelle
- Autres : _____

Merci beaucoup!

Annexe 7. Guide pour les entrevues individuelles
Personnel enseignant et membres de comités d'évaluation des apprentissages

PRÉPARATION

- Disposition de la salle
- Détermination de la place de l'animatrice
- Installation et vérification du matériel d'enregistrement

- Matériel nécessaire :
 - Enregistreuse;
 - Papier et crayons pour la prise de notes;
 - Formulaire de consentement;
 - Questionnaire sociodémographique;
 - Documents résumés (analyse de discours et analyse de concept);
 - Guide d'entrevue.

ACCUEIL

- Remise du formulaire de consentement et du questionnaire sociodémographique

PRÉSENTATION ET REMERCIEMENTS

- Présentations
- Remerciements
 - Tout d’abord, nous vous remercions d’avoir accepté de nous rencontrer malgré vos horaires chargés.
 - Dans cette deuxième partie de la recherche, les données seront recueillies par 1) des entrevues individuelles et des groupes de discussion focalisée avec des membres du personnel enseignant (utilisateurs) provenant des universités de Sherbrooke, McGill, Montréal et Laval et 2) des entrevues individuelles avec des acteurs influents au Canada.
 - Comme notre compréhension actuelle de cette conceptualisation est fondée principalement sur l'analyse de documents, votre participation au groupe de discussion focalisée nous permettra d’explorer l’acceptabilité des caractéristiques du concept de **la validité en tant qu’impératif social dans un contexte d’évaluation des apprentissages** auprès des personnes qui utilisent, développent, monitorent ou conseillent le développement et réfléchissent à propos des stratégies d’évaluation des apprentissages.
 - En tant que membre du personnel enseignant, vous jouez un rôle important dans l’évaluation des apprentissages des futurs professionnels de la santé.
 - Vous êtes bien placés pour nous informer et nous expliquer votre avis sur l’acceptabilité du concept émergent de la validité en tant qu’impératif social. Je vous rappelle que votre opinion est précieuse pour nous.

DÉROULEMENT DE LA RENCONTRE

- Présentation du déroulement plan de la rencontre :
 1. Complétion du formulaire de consentement et du questionnaire sociodémographique
Nous vous laisserons quelques minutes pour compléter les documents.
 2. Présentation des résultats de l'analyse de discours et de l'analyse de concept.
Vous avez reçu un résumé par courriel mais nous prendrons quelques minutes pour vous expliquer le projet et les résultats préliminaires.
 3. Entrevue individuelle
 4. Conclusion et validation des résultats
Vous recevrez par courriel un résumé des résultats préliminaires découlant de l'analyse des données.
- Explication du rôle de l'animatrice et de la personne participante
 - Animatrice :
 - Pose les questions visant à recueillir le point de vue.
 - Personnes participantes :
 - Partagent leur point de vue, leur perception et leur expérience avec l'animatrice.

CONFIDENTIALITÉ ET ANONYMAT

- Lire le formulaire de consentement à haute voix avec le groupe.
- Donner des explications supplémentaires au besoin.
- Vérifier régulièrement si la personne participante a des questions, si elle comprend bien le contenu du formulaire.

- Quelques éléments à aborder avec la personne participante :
 - Objectif poursuivi et thèmes traités pendant la rencontre;
 - Enregistrement évite la prise de notes, servira seulement aux chercheurs aux fins de la transcription et de l'analyse;
 - Contenu des échanges sera traité de manière à respecter l'anonymat des personnes participantes (données dénominalisées et identification par un code numérique).

- À la lumière de ces renseignements, si la personne participante accepte de participer à l'étude, lui demander de signer le formulaire de consentement en deux copies et lui en laisser une, l'autre sera déposée sous clé à l'endroit désigné.

- Faire remplir le questionnaire sociodémographique.

ENTREVUE

- Au cours de la présente rencontre, je vous poserai des questions afin de recueillir votre impression générale et ensuite votre point de vue concernant l'acceptabilité des attributs du concept émergent.
- La rencontre comportera au total 4 questions principales.
- Il n'y a pas de bonne ou de mauvaise réponse aux questions.

GUIDE D'ENTREVUE POUR LES ENTREVUES INDIVIDUELLES

Question 1 – Ouverture / brise-glace

Quel est votre nom et quel est votre rôle à l'Université [nom]?

Question 2- Concept général de la validité

Parlez-moi de votre vision du processus de validation (processus qui mesure la qualité) des outils d'évaluation des apprentissages dans votre cours et/ou programme?

Thèmes à explorer, si non abordés spontanément :

- Quelle importance accordez-vous à la mesure de la qualité des stratégies d'évaluation des apprentissages dans votre cours et/ou programme?

Question 3 –Attributs de l'analyse de concept

Que pensez-vous des caractéristiques que nous avons identifiées pour décrire une nouvelle manière d'assurer la qualité de l'évaluation des apprentissages dans une perspective de responsabilité sociale?

Thèmes à explorer, si non abordés spontanément :

- Quelle est la première chose qui vous vient à l'esprit quand vous lisez les attributs (caractéristiques) du concept de la validité en tant qu'impératif social?
- Que pensez-vous des attributs suivants?

Démonstration du recours à des preuves crédibles aux yeux de la société pour documenter la qualité de l'évaluation des apprenants

Démonstration d'une préoccupation constante des conséquences potentielles découlant de l'interprétation des résultats

Démonstration :

1. Des choix justifiés et raisonnés quant aux différentes stratégies d'évaluation des apprenants (leurs objectifs et fréquences, les contenus mesurés et leur imbrication)
2. De l'interprétation de la globalité de l'évaluation (dans son ensemble) et non séparément

Démonstration d'une utilisation justifiée de preuves mixtes (quantitatives et qualitatives) pour documenter la qualité de l'ensemble des stratégies de l'évaluation des apprentissages.

Question 4- Acceptabilité du concept émergent

Quelle est votre opinion personnelle face à l'acceptabilité du concept émergent de la validité en tant qu'impératif social dans votre cours/programme?

Thèmes à explorer, si non abordés spontanément :

- Pertinence du problème
- Commodité (applicabilité dans la réalité)
- Efficacité (utilité de l'intervention pour répondre au problème)
- Risques
- Adhésion (prêt à suivre les recommandations ou non)

CONCLUSION

- Remerciement de la présence de la personne participante
- Toute l'information transmise pourra éclairer les chercheuses et aider la réflexion sur la description du concept et son acceptabilité.

Annexe 8. Guide pour les groupes de discussion focalisée
Personnel enseignant et membres de comités d'évaluation des apprentissages

PRÉPARATION

- Disposition de la salle (tables, chaises disposées de manière à ce que les personnes participantes soient en cercle et qu'elles aient suffisamment d'espace pour leurs documents)
- Détermination de la place de l'animatrice et de l'observatrice
- Installation et vérification du matériel d'enregistrement

- Matériel nécessaire :
 - Enregistreuse;
 - Papier et crayons pour la prise de notes ;
 - Cartons pour écrire le nom des participants;
 - Formulaire de consentement;
 - Questionnaire sociodémographique;
 - Documents résumés (analyse de discours et analyse de concept);
 - Guide d'entrevue;
 - Feuille de présence;
 - Boissons et collation.

ACCUEIL

- Souhaiter la bienvenue aux personnes participantes.
- Remettre le carton avec leur nom (à placer devant eux).
- Signature de la feuille de présences.
- Inviter à se servir une boisson ou autre.
- Remise du formulaire de consentement et du questionnaire sociodémographique.

PRÉSENTATION ET REMERCIEMENTS

- Présentations/tour de table
- Remerciements
 - Tout d’abord, nous vous remercions d’avoir accepté de nous rencontrer malgré vos horaires chargés.
 - Dans cette deuxième partie de la recherche, les données seront recueillies par 1) des groupes de discussion focalisée avec des membres du personnel enseignant (utilisateurs) provenant des universités de Sherbrooke, McGill, Montréal et Laval et 2) des entrevues individuelles avec des acteurs influents au Canada.
 - Comme notre compréhension actuelle de cette conceptualisation est fondée principalement sur l'analyse de documents, votre participation au groupe de discussion focalisée nous permettra d’explorer l’acceptabilité des caractéristiques du concept de **la validité en tant qu’impératif social dans un contexte d’évaluation des apprentissages** auprès des personnes qui utilisent, développent, monitorent ou conseillent le développement et réfléchissent à propos des stratégies d’évaluation des apprentissages.
 - En tant que membre du personnel enseignant, vous jouez un rôle important dans l’évaluation des apprentissages des futurs professionnels de la santé.
 - Vous êtes bien placés pour nous informer et nous expliquer votre avis sur l’acceptabilité du concept émergent de la validité en tant qu’impératif social. Je vous rappelle que votre opinion est précieuse pour nous.

DÉROULEMENT DE LA RENCONTRE

- Présentation du déroulement plan de la rencontre :
- 5. Complétion du formulaire de consentement et du questionnaire sociodémographique
Nous vous laisserons quelques minutes pour compléter les documents.
- 6. Présentation des résultats de l'analyse de discours et de l'analyse de concept.
Vous avez reçu un résumé par courriel mais nous prendrons quelques minutes pour vous expliquer le projet et les résultats préliminaires.
- 7. Groupe de discussion focalisée
- 8. Conclusion et validation des résultats
Vous recevrez par courriel un résumé des résultats préliminaires découlant de l'analyse des données.
- Explication du rôle de l'animatrice, de l'observatrice et des personnes participantes.
 - Animatrice :
 - Introduit le groupe de discussion focalisée.
 - Pose les questions visant à recueillir le point de vue.
 - Encourage la participation de chacune des personnes.
 - Observatrice :
 - Prend des notes sur le déroulement de la rencontre.
 - Vérifie si les éléments importants ont tous été abordés.
 - Peut proposer des questions d'approfondissement au besoin.
 - Personnes participantes :
 - Partagent leur point de vue, leur perception et leur expérience avec les autres membres du groupe.
 - Échangent avec les autres, à la suite du partage des divers points de vue.

CONFIDENTIALITÉ ET ANONYMAT

- Lire le formulaire de consentement à haute voix avec le groupe.
- Donner des explications supplémentaires au besoin.
- Vérifier régulièrement si les participants ont des questions, s'ils comprennent bien le contenu du formulaire.

- Quelques éléments à aborder avec le groupe :
 - Objectif poursuivi et thèmes traités pendant la rencontre;
 - Enregistrement évite la prise de notes, servira seulement aux chercheurs aux fins de la transcription et de l'analyse;
 - Contenu des échanges sera traité de manière à respecter l'anonymat des personnes participantes (données dénominalisées et identification par un code numérique);
 - Contenu de la discussion de la rencontre de groupe est confidentiel. Chaque participant s'engage à respecter la confidentialité des échanges et à ne pas divulguer le nom des autres participants.

- Préciser aux participants qu'ils peuvent se retirer de la discussion de groupe en tout temps, qu'ils peuvent demander d'en effacer certains extraits sans que cela ne leur cause de préjudice.

- À la lumière de ces renseignements, si les participants acceptent de participer à l'étude, leur demander de signer le formulaire de consentement en deux copies et leur en laisser une, l'autre sera déposée sous clé à l'endroit désigné.

- Faire remplir le questionnaire sociodémographique.

ENTREVUE

- Au cours de la présente rencontre, je vous poserai des questions afin de recueillir votre impression générale et ensuite votre point de vue concernant l'acceptabilité des attributs du concept émergent.
- La rencontre comportera au total 4 questions principales.
- Il n'y a pas de bonne ou de mauvaise réponse aux questions.

GUIDE D'ENTREVUE POUR LES GROUPES DE DISCUSSION FOCALISÉE

Question 1 – Ouverture / brise-glace

Quel est votre nom et quel est votre rôle à l'Université de [nom]?

Question 2- Concept général de la validité

Parlez-moi de votre vision du processus de validation (processus qui mesure la qualité) des outils d'évaluation des apprentissages dans votre cours et/ou programme?

Thèmes à explorer, si non abordés spontanément :

- Quelle importance accordez-vous à la mesure de la qualité des stratégies d'évaluation des apprentissages dans votre cours et/ou programme?

Question 3 –Attributs de l’analyse de concept

Que pensez-vous des caractéristiques que nous avons identifiées pour décrire une nouvelle manière d’assurer la qualité de l’évaluation des apprentissages dans une perspective de responsabilité sociale?

Thèmes à explorer, si non abordés spontanément :

- Quelle est la première chose qui vous vient à l’esprit quand vous lisez les attributs (caractéristiques) du concept de la validité en tant qu’impératif social?
- Que pensez-vous des attributs suivants?

Démonstration du recours à des preuves crédibles aux yeux de la société pour documenter la qualité de l’évaluation des apprenants

Démonstration d’une préoccupation constante des conséquences potentielles découlant de l’interprétation des résultats.

Démonstration :

3. Des choix justifiés et raisonnés quant aux différentes stratégies d’évaluation des apprenants (leurs objectifs et fréquences, les contenus mesurés et leur imbrication)
4. De l’interprétation de la globalité de l’évaluation (dans son ensemble) et non séparément

Démonstration d’une utilisation justifiée de preuves mixtes (quantitatives et qualitatives) pour documenter la qualité de l’ensemble des stratégies de l’évaluation des apprentissages.

Question 4- Acceptabilité du concept émergent

Quelle est votre opinion personnelle face à l’acceptabilité du concept émergent de la validité en tant qu’impératif social dans votre cours/programme?

Thèmes à explorer, si non abordés spontanément :

- Pertinence du problème
- Commodité (applicabilité dans la réalité)
- Efficacité (utilité de l’intervention pour répondre au problème)
- Risques
- Adhésion (prêt à suivre les recommandations ou non)

CONCLUSION

- Remerciement de la présence des participants.
- Toute l'information transmise pourra éclairer les chercheuses et aider la réflexion sur la description du concept et son acceptabilité.

Annexe 9. Résumé de l'analyse de discours (version française)

**Validité : Un mot avec de multiples significations
en pédagogie des sciences de la santé**

Christina St-Onge, Meredith Young, Kevin Eva, & Brian Hodges

Référence :

St-Onge, C., Young, M., Eva, K., Hodges, B. (2016, on-line first). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education*, DOI: 10.1007/s10459-016-9716-3

Selon certaines revues systématiques réalisées au cours des dernières années, il semble que les pratiques de validation retrouvées dans la littérature en pédagogie des sciences de la santé, diffèrent des recommandations à ce sujet. Il semble nécessaire de mieux comprendre ces différences entre la théorie et la pratique puisque le concept de la validité est essentiel pour le développement et le monitoring de stratégies et de programmes d'évaluation. Or, l'objectif de ce projet était d'identifier les différentes conceptualisations de la validité dans les écrits scientifiques en pédagogie des sciences de la santé.

Nous avons réalisé une analyse de discours afin de documenter les multiples façons de conceptualiser la validité de l'évaluation des apprentissages et des compétences en pédagogie des sciences de la santé à l'aide d'un échantillon stratégique qui incluait les références clés utilisées par les co-chercheurs, bonifiées par une recension des écrits effectués avec les bases de données PubMed, ERIC, et PsychInfo. Un processus d'analyse itératif a été réalisé jusqu'à l'obtention d'un consensus pour identifier et définir les discours émergents.

Trois discours ont été identifiés. La *validité en tant que caractéristique d'un test* suppose que la validité peut être une propriété intrinsèque d'un instrument de mesure. La *validité en tant que processus d'accumulation d'évidences* met l'emphase sur l'importance de supporter, de façon adéquate et appropriée, les décisions prises à partir de résultats d'évaluation. La *validité en tant que responsabilité sociale* voit la validité comme une qualité au cœur des programmes d'évaluation mise de l'avant pour répondre aux besoins de la société et des apprenants. Les caractéristiques de ces discours sont résumées au

Tableau 1. Il appert que la validité peut prendre différentes conceptualisations selon le discours auquel on adhère. Il va sans dire que la validité est un terme omniprésent en évaluation et que les divergences observées (entre théorie et pratiques) peuvent possiblement être expliquées par des divergences de conceptualisations.

TABLEAU 1: Résumé des trois discours

Note: Nous avons conservé le texte intégral en anglais pour ne pas modifier le sens des énoncés

	Validity as a Test characteristic	Validity as an argument-based evidentiary-chain	Validity as a Social imperative
<i>Definition</i>	The degree to which the test actually measures what it purports to measure.	The evidences presented to support or refute the meaning or interpretation assigned to assessment results.	A bird's eye view of assessment that foregrounds broader individual and societal issues
<i>Characteristics</i>	Validity is a goal or a gold seal of approval.	Validity is a journey on which one embarks to provide evidence supporting the interpretation of scores.	Validity and validation are matters of social accountability.
<i>Validity is viewed as...</i>	Static	Fluid	Built-in
<i>Focus of evidence is on...</i>	Individual tools can be considered valid, and the validity can generalize to the tool format (MCQs are valid)	Defensible interpretation of scores	Individual and societal impact of assessment
<i>Things made possible</i>	The quest for the holy grail of assessment; one tool that is more valid than the others.	Validation approaches and standards	Holistic and <i>a priori</i> consideration for societal impact of assessment
<i>Validation occurs...</i>	A posteriori (mainly)		A priori (mainly)
<i>Validation data focused on...</i>	Psychometric	Mostly psychometric	Mostly expert judgment

Annexe 10. Résumé de l'analyse de concept (version française)

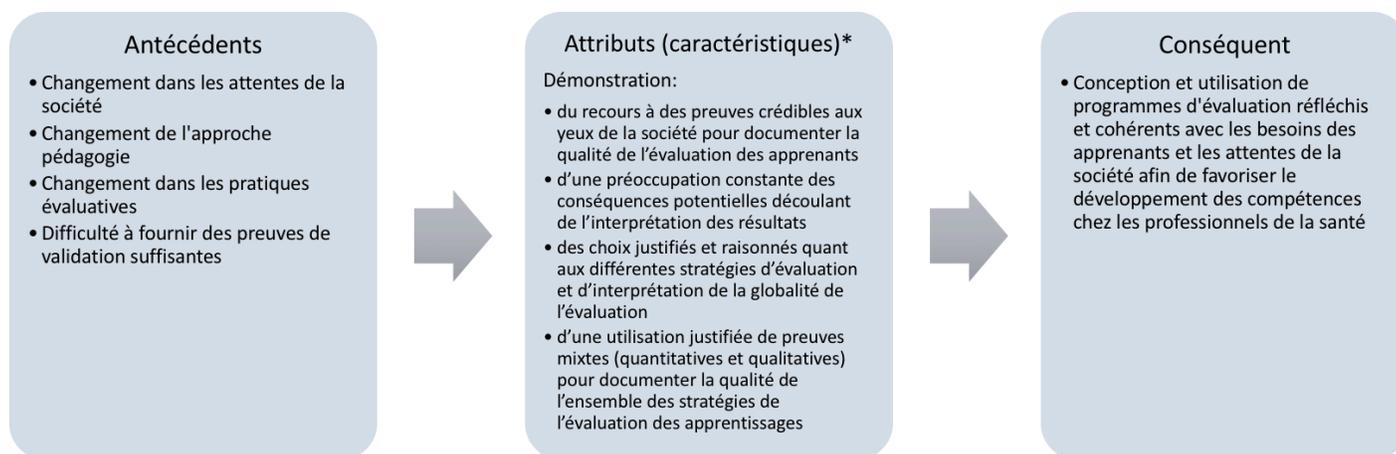
Contexte :

Depuis l'adoption de l'approche par compétences (APC) dans les programmes de formation en sciences de la santé, les changements touchant l'évaluation des apprentissages ont été multiples tels que l'utilisation de l'évaluation authentique (qui se rapproche de la pratique professionnelle réelle) ou de la combinaison de différentes stratégies d'évaluation, contextes et temps de mesure. Comme les résultats découlant de l'évaluation des apprentissages peuvent avoir des conséquences importantes pour le futur professionnel de la santé et pour la société, la mesure de la qualité (validité) des stratégies d'évaluation est essentielle (Van der Vleuten & Schuwirth, 2005). Toutefois, un écart existe entre les méthodes de validation présentement à notre disposition et la réalité vécue par les responsables de l'évaluation (Van der Vleuten & Schuwirth, 2005). Cet écart a mené à des réflexions qui ont été résumées dans une récente analyse de discours (St-Onge et al., 2017). Parmi les résultats obtenus par cette analyse de discours, on retrouve une nouvelle conceptualisation de la validité : **la validité en tant qu'impératif social dans un contexte d'évaluation des apprentissages.**

Analyse de concept :

Déjà terminée, la première étape de la recherche doctorale consistait à décrire le concept émergent à l'aide d'une analyse de concept selon la méthode de Rodgers (1989). L'analyse de concept est une méthode utile pour identifier, clarifier et raffiner un concept inexploré (Rodgers, 1989; Rodgers & Knafl, 2000). Cette étape a représenté une étape cruciale dans le processus du développement des connaissances.

Résultats de l'analyse de concept



*Les attributs (caractéristiques) sont décrits en profondeur à la page suivante

Et maintenant, c'est à votre tour!

Comme notre compréhension actuelle de cette conceptualisation est fondée principalement sur l'analyse de documents, votre participation au groupe de discussion focalisée nous permettra d'explorer l'acceptabilité des caractéristiques du concept de **la validité en tant qu'impératif social dans un contexte d'évaluation des apprentissages** auprès des personnes qui utilisent, développent, monitorent ou conseillent le développement et réfléchissent à propos des stratégies d'évaluation des apprentissages.

Attributs (caractéristiques)	Description
<p>Démonstration du recours à des preuves crédibles aux yeux de la société pour documenter la qualité de l'évaluation des apprenants.</p>	<p>Les différentes instances professionnelles (établissement d'enseignement et ordres professionnels) doivent pouvoir documenter avec certitude - pour la société - les décisions prises quant au cheminement des étudiants et leur niveau d'aptitudes à débiter une carrière professionnelle de façon autonome et compétente.</p> <p>Par exemple : Un établissement universitaire est imputable auprès de la société quant aux décisions prises à partir de l'évaluation des apprenants.</p> <p>« <i>When students graduate from a university, the degree indicates to society that the graduates have a certain level of skill and expertise</i> (Boley & Whitney, 2003, p.199) ».</p>
<p>Démonstration d'une préoccupation constante des conséquences potentielles découlant de l'interprétation des résultats.</p>	<p>La justification des décisions prises durant le processus d'élaboration, d'administration et d'interprétation des résultats repose sur des considérations des conséquences potentielles que l'interprétation des résultats d'évaluation pourrait entraîner sur l'individu, l'institution, et la société.</p> <p>En plus de mesurer les conséquences découlant de l'interprétation des résultats de l'évaluation des apprentissages, la personne responsable des évaluations doit anticiper les conséquences potentielles liées à cette interprétation et mettre en place des moyens ou des stratégies pour les minimiser. Les conséquences mesurées ne se limitent pas seulement aux impacts liés au construit, mais concernent l'ensemble des conséquences possibles.</p> <p>« (...) <i>the measurement or scoring procedure (e.g., irrelevant, unreliable, or omitted test items); the specific interpretation (e.g., an inappropriate pass/fail cut point); the attribute being measured (i.e., the wrong construct); or the response (e.g., the actions that follow the decision)</i>. (Cook & Lineberry, 2016, p.790) »</p>
<p>Démonstration</p> <ol style="list-style-type: none"> 1- Des choix justifiés et raisonnés quant aux différentes stratégies d'évaluation des apprenants (leurs objectifs et fréquences, les contenus mesurés et leur imbrication) 2- De l'interprétation de la globalité de l'évaluation (dans son ensemble) et non séparément 	<p>La planification et l'élaboration de l'évaluation (p. ex. : choix des stratégies, objectifs, fréquences, contenus), l'administration ainsi que l'interprétation des résultats de l'évaluation des apprenants doit être réalisée de façon à considérer un « tout » (le programme d'évaluation dans son entier) qui est supérieur à la somme de ses parties.</p> <p>« <i>The central key is that the programme of assessment is set up to allow the whole picture of a student's competence to be obtained by a careful selection of assessment methods, formulation of rules and regulations and design of organisational systems</i> (Schuwirth & Van der Vleuten, 2011, p.479) ».</p>
<p>Démonstration d'une utilisation justifiée de preuves mixtes (quantitatives et qualitatives) pour documenter la qualité de l'ensemble des stratégies de l'évaluation des apprentissages.</p>	<p>Comme les stratégies quantitatives traditionnelles (p. ex. : Alpha de Cronbach) ne sont pas appropriées pour démontrer la qualité d'un ensemble des stratégies d'évaluation des apprentissages (p. ex. : programme d'évaluation), la combinaison de preuves de types quantitatives et qualitatives semble être une option intéressante.</p> <p>« <i>In our efforts to proceed beyond the psychometric discourse, we find inspiration in methodologies from qualitative research. As in quantitative research, rigour is built into qualitative research (...) For each of these notions, methodological strategies are proposed that bring rigour to the research: prolonged engagement, triangulation, peer examination, member checking, structural coherence, time sampling, stepwise replication, audit and thick description. With some creativity, we can apply these strategies to assessment to achieve rigour of decision making</i> (Van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010, p.713) ».</p>

**Annexe 11. Formulaire de consentement pour les entrevues individuelles
Personnel enseignant et membres de comités d'évaluation des apprentissages**

FORMULAIRE DE CONSENTEMENT POUR LE PROJET DE RECHERCHE

**La validité en tant qu'impératif social
dans le contexte de l'évaluation des apprentissages**

Mélanie Marceau, inf. M.Sc.
Candidate au doctorat en sciences cliniques (pédagogie des sciences de la santé)
Université de Sherbrooke

Équipe de direction :
Frances Gallagher¹, Ph.D., Christina St-Onge¹, Ph.D., Meredith Young², Ph.D.,
¹Faculté de médecine et sciences de la santé, Université de Sherbrooke
²Faculté de médecine, Université McGill

Madame,
Monsieur,

Vous avez été invités à participer à la recherche en titre. Le but de cette étude sera d'explorer le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages. Ce projet a été approuvé par le Comité d'éthique de la recherche de l'Université [nom et numéro d'approbation].

En quoi consiste la participation au projet?

Votre contribution à ce projet de recherche consiste à participer à une entrevue aujourd'hui le [date]. Cette entrevue se déroule en présentiel dans votre université. Le seul inconvénient lié à votre participation est le temps consacré à la recherche, soit environ 90 minutes.

À la suite de l'entrevue, l'équipe de recherche communiquera avec vous par courriel afin de partager une synthèse des résultats provenant des groupes de discussion focalisée ainsi que des entrevues individuelles afin de valider la compréhension.

Qu'est-ce que la chercheuse fera avec les données recueillies?

Pour éviter votre identification comme participant dans ce projet, les données seront traitées de manière entièrement confidentielle. Votre nom sera remplacé par un code numérique qui sera connu seulement de l'équipe de recherche et qui sera maintenu sous clé. Le système de codage sera détruit à la fin de l'étude. Les jugements et opinions partagées sur les questionnaires demeureront strictement anonymes. Cinq ans après la publication des résultats, les documents papier et électroniques seront détruits. Les données ne seront pas utilisées à d'autres fins que celles décrites dans le présent document : thèse, articles ou présentations scientifiques. Comme il s'agit d'une étude de nature exploratoire, l'équipe de recherche pourrait avoir besoin de retourner aux données recueillies pour approfondir l'analyse. À tout moment, les données demeureront anonymes et confidentielles.

Est-il obligatoire de participer?

Non. La participation à cette étude se fait sur une base volontaire. Vous êtes **entièrement libre de participer ou non**, et de vous retirer en tout temps sans avoir à motiver votre décision ni à subir de préjudice de quelque nature que ce soit.

Y a-t-il des risques, inconvénients ou bénéfices?

Au-delà des risques et inconvénients mentionnés jusqu'ici (temps consacré à la recherche), les chercheuses et chercheurs considèrent que les risques possibles sont minimaux. La contribution à l'avancement des connaissances au sujet de la validité en tant qu'impératif social sont les bénéfices prévus. Aucune compensation d'ordre monétaire n'est offerte pour la participation à ce projet. **Que faire si j'ai des questions concernant le projet?**

Toute plainte ou critique sur ce projet de recherche pourra être adressée au Bureau de l'Ombudsman de l'Université [nom et coordonnées] :

Si vous avez des questions concernant ce projet de recherche, n'hésitez pas à communiquer avec moi aux coordonnées indiquées ci-dessous.

Mélanie Marceau, inf. M.Sc.

Candidate au doctorat en sciences cliniques (pédagogie des sciences de la santé)

Faculté de médecine et des sciences de la santé

Université de Sherbrooke

[XXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

J'ai lu et compris le document d'information au sujet du projet Validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages. J'ai compris les conditions, les risques et les bienfaits de ma participation. J'ai obtenu des réponses aux questions que je me posais au sujet de ce projet. J'accepte librement de participer à ce projet de recherche.

J'accepte de participer à une entrevue

Participant(e) ou participant :	Membre de l'équipe de recherche
Signature :	Signature :

Nom :	Nom :
-------	-------

Date :	Date :
--------	--------

**S.V.P., signez les deux copies.
Conservez une copie et remettez l'autre à la chercheuse.**

Ce projet a été revu et approuvé par le comité d'éthique de la recherche Éducation et sciences sociales, de l'Université de Sherbrooke. Cette démarche vise à assurer la protection des participantes et participants. Si vous avez des questions sur les aspects éthiques de ce projet (consentement à participer, confidentialité, etc.), n'hésitez pas à communiquer avec M. Eric Yergeau, président de ce comité, par l'intermédiaire de son secrétariat aux coordonnées :

Pour Sherbrooke : XXX-XXX-XXXX poste XXXXX

Pour l'extérieur de Sherbrooke : (X XXX XXX-XXXX) poste XXXXX

Par courriel : XXXXXXXX

**Annexe 12. Formulaire de consentement pour les groupes de discussion focalisée
Personnel enseignant et membres de comités d'évaluation des apprentissages**

FORMULAIRE DE CONSENTEMENT POUR LE PROJET DE RECHERCHE

**La validité en tant qu'impératif social
dans le contexte de l'évaluation des apprentissages**

Mélanie Marceau, inf. M.Sc.
Candidate au doctorat en sciences cliniques (pédagogie des sciences de la santé)
Université de Sherbrooke

Équipe de direction :
Frances Gallagher¹, Ph.D., Christina St-Onge¹, Ph.D., Meredith Young², Ph.D.,
¹Faculté de médecine et sciences de la santé, Université de Sherbrooke
²Faculté de médecine, Université McGill

Madame,
Monsieur,

Vous avez été invités à participer à la recherche en titre. Le but de cette étude sera d'explorer le concept de validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages. Ce projet a été approuvé par le Comité d'éthique de la recherche de l'Université [nom et numéro d'approbation].

En quoi consiste la participation au projet ?

Votre contribution à ce projet de recherche consiste à participer à un groupe de discussion focalisée aujourd'hui le "indiquer date". Cette entrevue de groupe se déroule en présentiel dans votre université. Le seul inconvénient lié à votre participation est le temps consacré à la recherche, soit environ 90 minutes.

Suite au groupe de discussion focalisée, l'équipe de recherche communiquera avec vous par courriel afin de partager une synthèse de la rencontre et les résultats provenant de tous les groupes de discussion focalisée afin de valider la compréhension.

Qu'est-ce que la chercheuse fera avec les données recueillies?

Pour éviter votre identification comme participant dans ce projet, les données seront traitées de manière entièrement confidentielle. Votre nom sera remplacé par un code numérique qui sera connu seulement de l'équipe de recherche et qui sera maintenu sous clé. Le système de codage sera détruit à la fin de l'étude. Les jugements et opinions partagés sur les questionnaires demeureront strictement anonymes. Cinq ans après la publication des

résultats, les documents papier et électroniques seront détruits. Les données ne seront pas utilisées à d'autres fins que celle décrite dans le présent document : thèse, articles ou présentations scientifiques. Comme il s'agit d'une étude de nature exploratoire, l'équipe de recherche pourrait avoir besoin de retourner aux données recueillies pour approfondir l'analyse. À tout moment, les données demeureront anonymes et confidentielles.

Vous-mêmes et les autres personnes qui participeront au groupe de discussion focalisée connaîtrez l'identité des personnes participantes et des renseignements et opinions partagés lors de la discussion. Nous comptons sur votre collaboration afin de préserver la confidentialité de ces informations.

Est-il obligatoire de participer?

Non. La participation à cette étude se fait sur une base volontaire. Vous êtes **entièrement libre de participer ou non**, et de vous retirer en tout temps sans avoir à motiver votre décision ni à subir de préjudice de quelque nature que ce soit.

Y a-t-il des risques, inconvénients ou bénéfices?

Au-delà des risques et inconvénients mentionnés jusqu'ici (temps consacré à la recherche), les chercheuses et chercheurs considèrent que les risques possibles sont minimaux. La contribution à l'avancement des connaissances au sujet de la validité en tant qu'impératif social sont les bénéfices prévus. Aucune compensation d'ordre monétaire n'est offerte pour la participation à ce projet.

Que faire si j'ai des questions concernant le projet?

Si vous avez des questions concernant ce projet de recherche, n'hésitez pas à communiquer avec moi aux coordonnées indiquées ci-dessous.

Mélanie Marceau, inf. M.Sc.

Candidate au doctorat en sciences cliniques (pédagogie des sciences de la santé)

Faculté de médecine et des sciences de la santé

Université de Sherbrooke

[XXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

J'ai lu et compris le document d'information au sujet du projet Validité en tant qu'impératif social dans le contexte de l'évaluation des apprentissages. J'ai compris les conditions, les risques et les bienfaits de ma participation. J'ai obtenu des réponses aux questions que je me posais au sujet de ce projet. J'accepte librement de participer à ce projet de recherche.

- J'accepte de participer à un groupe de discussion focalisée*
- Je m'engage à respecter la confidentialité des renseignements partagés lors de la discussion de groupe focalisée (noms des autres participants et informations dévoilées)*

Participant(e) ou participant :	Membre de l'équipe de recherche
---------------------------------	---------------------------------

Signature :	Signature :
-------------	-------------

Nom :	Nom :
-------	-------

Date :	Date :
--------	--------

**S.V.P., signez les deux copies.
Conservez une copie et remettez l'autre à la chercheuse.**

Ce projet a été revu et approuvé par le comité d'éthique de la recherche Éducation et sciences sociales, de l'Université de Sherbrooke. Cette démarche vise à assurer la protection des participantes et participants. Si vous avez des questions sur les aspects éthiques de ce projet (consentement à participer, confidentialité, etc.), n'hésitez pas à communiquer avec M. Eric Yergeau, président de ce comité, par l'intermédiaire de son secrétariat aux coordonnées :

Pour Sherbrooke : XXX-XXX-XXXX poste XXXXX

Pour l'extérieur de Sherbrooke : (X XXX XXX-XXXX) poste XXXXX

Par courriel : XXXXXXXX

Annexe 13. Invitation à participer
Personnes engagées dans la recherche en pédagogie des sciences de la santé

INVITATION TO PARTICIPATE IN THE RESEARCH STUDY

Validity as a social imperative in assessment

Mélanie Marceau, R.N. M.Sc. (nursing)
Ph.D. candidate in clinical sciences (health sciences education)
Université de Sherbrooke

Supervisors:

Christina St-Onge¹, Ph.D., Meredith Young², Ph.D., Frances Gallagher¹, Ph.D.

¹Faculty of medicine and health sciences, Université de Sherbrooke

²Faculty of Medicine, McGill University

Dear [name],

You are being asked to participate in a research study investigating validity as a social imperative within the context of assessment in Health Professions Education. This study was approved by the research ethics board at Université de Sherbrooke, McGill University, Université Laval and Université de Montréal.

You have been identified as a potential participant because of your involvement in the field of assessment and validity in health professions education (HPE). Because of your contribution (research or peer-reviewed publications) and your work relevant to validity and assessment within the HPE literature, we feel that based on your expertise, you are well placed join us for a conversation on the acceptability of validity as a social imperative in HPE assessment.

In parallel to inviting you to participate in interviews, we are also conducting concurrent focus group with knowledge users (faculty members, course instructors, and teaching assistants) within university health programs in Québec (occupational therapy, physiotherapy, medicine, and nursing). Through this process, we hope to further our understanding of validity as a social imperative, and further explore whether this conceptualization resonates with individuals within HPE.

What is validity as a social imperative?

To provide more context to the current study, a recent discourse analysis done by the members of our research team revealed dissatisfaction with the available validation strategies (St-Onge et al., 2017), as well as a new conceptualization of validity, namely **validity as a social imperative**.

Since our current understanding of validity as a social imperative is based mainly on literature reviews, this emerging concept requires much more in-depth study, in particular, to determine its acceptability to the people directly involved in the learning assessment, and to individuals within the assessment and HPE communities.

What will my participation involve?

Your contribution to this research study consists of participating in online (skype) or face to face (when available) interview. We anticipate the interviews to take no more than 90 minutes, and can be conducted in French or English, according to your preference.

If you agree to participate in an interview, prior to our conversation, you will receive an email regarding the context of the research study, with some material defining validity as a social imperative (discourse analysis and concept analysis). Prior to the interview, you will be asked to complete and transmit a consent form and a sociodemographic questionnaire.

If you would like to participate in this study, please contact our research assistant (Kathleen Day), and she will arrange a mutually beneficial time for the interviews.

Following the interview, the research team will email you a summary of the results to validate the conclusions of the data analysis.

What will the researcher do with the data collected?

To prevent you being identified as a participant in this study, all data will be kept confidential. Your name will be replaced with a numeric code, known only to the research team and kept in a password protected computer in a locked office. The key to the coding system will be destroyed at the end of the study. The statements and opinions made on the questionnaires will remain strictly confidential. Five years after the publication of the results, the paper and electronic documents will be destroyed. The data will not be used for purposes other than those described in this document (theses, scientific articles, or presentations). Since this is an exploratory study, the research team may need to periodically consult the data collected in order to perform more extensive analyses. The data will remain confidential at all times.

Is my participation mandatory?

No. Your participation in this study is voluntary. You are **entirely free to participate or not**, and to withdraw from the study at any time, without having to explain your decision; withdrawing from the study will not result in any negative consequences for you.

Are there any risks, drawbacks, or benefits related to the study?

Apart from the time commitment of the interview, the researchers consider the possible risks to be minimal. The expected benefit is the contribution to the advancement of knowledge about validity as a social imperative. You will not be paid to participate in this study.

What if I have questions about the study?

Feel free to contact the undersigned if you have any questions about this research study.

Mélanie Marceau, R.N., M.Sc. (nursing)

Ph.D. candidate in clinical sciences (health sciences education)

Faculty of medicine and health sciences

Université de Sherbrooke

[XXXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

Kathleen Day, MA

Interviewer

[XXXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

Christina St-Onge, Ph.D.

Titulaire de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke

Associate professor, Department of Medicine,

Faculty of medicine and health sciences, Université de Sherbrooke

[XXXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

Meredith Young, Ph.D.

Assistant Professor

Centre for Medical Education

Faculty of Medicine, McGill University

[XXXXXXXXXX](#)

Tel. : (XXX) XXX-XXXX

Frances Gallagher, R.N., Ph.D.

Co-director of the undergraduate program

Associate professor

Nursing School

Faculty of medicine and health sciences, Université de Sherbrooke

[XXXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

Annexe 14. Questionnaire sociodémographique
Personnes engagées dans la recherche en pédagogie des sciences de la santé

Here is a short socio-demographic questionnaire.

Again, thank you very much for your participation in this project.

1. What is your gender:
 Male?
 Female?

2. What is your age?
 25-34 years old
 35-44 years old
 45-54 years old
 55-64 years old
 65 years or older

3. Which of the following categories best describes your initial training discipline?
 Nursing
 Medicine
 Occupational therapy
 Physical therapy
 Education
 Other: _____

4. In which field do you mainly work
 Nursing
 Medicine
 Occupational therapy
 Physical therapy
 Education
 Other: _____

5. How long have you been working in this function? _____years

Please indicate your level of agreement with each of the following statements:

1 = Totally disagree	2 = Disagree	3 = Somewhat disagree
4 = Somewhat agree	5 = Agree	6 = Totally agree

Please tell us a little bit about your experience about validity in the context of assessment

6. I regularly work with assessment concepts.	1	2	3	4	5	6
7. I often think about validity in terms of my work functions.	1	2	3	4	5	6
8. Most of my tasks are related to assessment issues.	1	2	3	4	5	6
9. I think I have a good understanding of validity issues.	1	2	3	4	5	6

Annexe 15. Formulaire de consentement
Personnes engagées dans la recherche en pédagogie des sciences de la santé

Validity as a social imperative in assessment

Mélanie Marceau, R.N. M.Sc. (nursing)
Ph.D. candidate in clinical sciences (health sciences education)
Université de Sherbrooke

Supervisors:

Christina St-Onge¹, Ph.D., Meredith Young², Ph.D., Frances Gallagher¹, Ph.D.

¹Faculty of medicine and health sciences, Université de Sherbrooke

²Faculty of Medicine, McGill University

Sir/Madam

You are being asked to participate in a research study investigating validity as a social imperative within the context of assessment in Health Professions Education (HBE). This study was approved by the research ethics board at Université de Sherbrooke, McGill University, Université Laval and Université de Montréal.

What will my participation involve?

Your contribution to this project consists of participating in an interview – either in person, over Skype, or by phone. The interview should take between 60 and 90 minutes. Following the interview, the research team will email you a summary of your interview and you will be asked to complete and validate it.

What will the researcher do with the data collected?

To prevent you being identified as a participant in this study, all data will be kept confidential. In the transcripts, your name will be replaced with a numeric code, known only to the research team and kept in a password protected computer in a locked office. The key to the coding system will be destroyed at the end of the study. The statements and opinions made on the questionnaires and within the context of the interview, will remain strictly confidential. Five years after the publication of the results, the paper and electronic documents will be destroyed. The data will not be used for purposes other than those described in this document (theses, scientific articles, or presentations). Since this is an exploratory study, the research team may need to periodically consult the data collected in order to perform more extensive analyses. The data will remain confidential at all times.

Is my participation mandatory?

No. Your participation in this study is voluntary. You are **entirely free to participate or not**, and to withdraw from the study at any time, without having to explain your decision. Withdrawing from the study will not result in any negative consequences for you.

Are there any risks, drawbacks, or benefits related to the study?

Apart from the time commitment of the interview, the researchers consider the possible risks to be minimal. The expected benefit is the contribution to the advancement of knowledge about validity as a social imperative. You will not be paid or compensated to participate in this study.

What if I have questions about the study?

Feel free to contact the undersigned if you have any questions about this research study.

Mélanie Marceau, R.N., M.Sc. (nursing)

Ph.D. candidate in clinical sciences (health sciences education)

Faculty of medicine and health sciences

Université de Sherbrooke

[XXXXXXXXX](#)

Tel.: XXX-XXX-XXXX, ext. XXXXX

I have read and understood the consent form introducing the study on Validity as a social imperative in the context of the assessment. I understood the conditions, the risks and the benefits of my participation. I have received answers to the questions I had about this project. I agree freely to participate in this research project.

I agree to participate in this project

Participant :	Member of the team research
Signature :	Signature :

Name :	Name :
--------	--------

Date :	Date :
--------	--------

**Please, sign and send a copy or click the following link [link]
to indicate your willingness to participate in the study.**

This project was reviewed and approved by the Research Ethics Committee for Education and Social Sciences, Université de Sherbrooke. This is to ensure the protection of the participants. If you have any questions about the ethical aspects of this project (consent to participate, confidentiality, etc.), please do not hesitate to contact Eric Yergeau, president of this committee, through its coordinating secretariat:

Sherbrooke: XXX-XXX-XXXX ext. XXXXX

Outside of Sherbrooke: (X XXX XXX-XXXX) ext. XXXXX

By e-mail: [XXXXXXXXX](#)

Annexe 16. Guide pour les entrevues individuelles Personnes engagées dans la recherche en pédagogie des sciences de la santé

Introduction

Presentations

First of all, thank you for accepting our invitation to participate in this study despite your busy schedules.

You have been identified as a potential participant because of your involvement in the field of validity and assessment in health professions' education (HPE). Because of your contribution (research or peer-reviewed publications) and your work relevant to assessment within the HPE literature, you are well placed to have a conversation with us around the emerging concept of validity as social imperative.

In the context of this conversation, there are no good or bad answers, we really are looking forward to your opinion.

Overview of the interview

Presentation of the expected overall plan for the interview

Reminder of consent form and demographic survey

Previously, you have completed the consent form and the demographic survey. If you have not had the time to fill them, please click the following link :

https://www.sondageonline.com/s/validity_social_imperative

to indicate your willingness to participate in the study before we start.

(Note for the interviewer: send back the link by email if needed and give some time to fill the consent form and the demographic survey)

Presentation of the results of the discourse analysis and concept analysis.

You have received a short email describing the current study with two documents attached to describe the previous stages of this program of work.

Specifically, the first document, a summary of a recent **discourse analysis**, explains how the concept of validity as a social imperative emerged. In short, an empirical form of discourse analysis documented several ways in which validity is described, understood, and used in health professions education.

The second document, the summary of our results from a **concept analysis**, is the central element of today's conversation. The emerging concept of validity as social imperative, to date, has been based mainly on a review of the literature. The purpose of this conversation

is to gauge the acceptability of the concept of validity as a social imperative. Because of your work relevant to assessment within HPE, your perspective will help to refine our concept.

Interview

Conclusion and validation of the summary of your interview.

Following data analysis, the research team will forward a summary of your interview which you can complete and validate if necessary.

Confidentiality and anonymity

Indications for the interviewer.

Reminder to be mentioned to participants before the beginning of the interview:

- The interview will be audio-recorded. The recording will only be used for research purposes, and will be transcribed.
- Content exchanges will be treated in such a way as to respect the anonymity of the participants (name of the data and identification by a numerical code).
- Specify that participants may withdraw from the interview at any time, and may request that certain excerpts be deleted without harm.
- Ask if they have any questions before starting the interview

Context

In order to explore the acceptability of the concept of validity as a social imperative (Phase 2), we are seeking feedback from two main groups. You have been invited to participate as an influential actor, paired with several focus groups with assessment developers and users from four universities (Sherbrooke, McGill, Montreal and Laval), and across several programs in health sciences (occupational therapy, physical therapy, nursing and medicine).

Interview

The interview will include a total of 5 main questions.

There is no right or wrong answer to the questions.

INTERVIEW GUIDE FOR INTERVIEW

Question 1 - Opening / icebreaker

Can you tell me what motivated you to participate in this study?

Question 2 - General concept of validity

Would you tell me what validity in the context of assessment means to you?

Deepening's question:

Can you explain your point of view?

What brings you to ...?

Can you give me an example?

Question 3 -Attributes of concept analysis

I would like you to take a look at the second document that we circulated (results of the concept analysis). What do you think of the attributes (characteristics) that we identified to describe validity as social imperative? (Discuss one by one)

Characteristics resulting from the concept analysis

Demonstration of the use of evidence considered credible by society to document the quality of assessments

Validation embedded through the assessment process and score interpretation

Documented validity evidence supporting the interpretation of the combination of assessment findings

Demonstration of a justified use of a variety of evidence (quantitative and qualitative) to document the quality of all assessment strategies.

Themes to explore, if not addressed spontaneously or if they want more precision about our definition of acceptability:

Relevance of the attribute

Convenience (applicability to practice)

Efficiency (use of the attribute to address the validity of assessment as a social imperative)

Risks

Adhesion (ready to follow the recommendations or not)

Deepening's question:

Can you explain your point of view?

What brings you to ...?

Can you give me an example?

What else can you say about that?
 What does X word mean to you?
 What are some ways that X is different than Y?
 What do you mean when you say X is ['no good']?
 What do you like about X?"
 What don't you like about X?

Question 4 Acceptability of emerging concept overall

Overall, what is your personal opinion of validity as social imperative?

Themes to explore, if not addressed spontaneously or if they want more precision about our definition of acceptability:

- Relevance of the problem
- Convenience (applicability in reality; or adaptability to practice)
- Efficiency (use of intervention to address the problem)
- Risks
- Adhesion (ready to follow the recommendations or not)

Deepening's question:

Can you explain your point of view?
 What brings you to ...?
 Can you give me an example?
 What else can you say about that?
 What does X word mean to you?
 What are some ways that X is different than Y?
 What do you mean when you say X is ['no good']?
 What do you like about X?"
 What don't you like about X?

Question 5 Conclusion

Do you feel that we have missed important aspects of validity as a social imperative?
 Do you have anything else you would like to add? Something else we should consider as we move this work forward?

Deepening's question:

What else can you say about that?
 Can you explain your point of view?
 Is there anything else you can add?
 Can you build on that?

Conclusion

Indications for the interviewer.

Thank the participant

Inform the participant that:

All submitted information will help inform researchers and reflection on the description of the concept and its acceptability.

Following data analysis, the research team will forward a summary of your interview which you can complete and validate if necessary.

It is possible that the research team will contact the participants again for Phase 4 of the research where we will be attempting to operationalize validity as a social imperative. More specifically, we will aim to document the acceptability and the feasibility of the operationalization of the concept (for example: comment on the clarity and relevance of the markers and indicators, the choice of format, etc.). You will be free to accept or not to participate in this subsequent phase.

Annexe 17. Résumé de l'analyse de discours (version anglaise)

Validity: One word with a plurality of meanings.

Christina St-Onge, Meredith Young, Kevin Eva, & Brian Hodges

Reference:

St-Onge, C., Young, M., Eva, K., Hodges, B. (2016, on-line first). Validity: One word with a plurality of meanings. *Advances in Health Sciences Education*, DOI: 10.1007/s10459-016-9716-3

Validity is one of the most debated constructs in our field; debates abound about what is legitimate and what is not, and the word continues to be used in ways that are explicitly disavowed by current practice guidelines. The resultant tensions have not been well characterized, yet their existence suggests that different uses may maintain some value for the user that needs to be better understood. We conducted an empirical form of Discourse

Analysis to document the multiple ways in which validity is described, understood, and used in the health professions education field. We created and analyzed an archive of texts identified from multiple sources, including formal databases such as PubMed, ERIC and PsycINFO as well as the authors' personal assessment libraries. An iterative analytic process was used to identify, discuss, and characterize emerging discourses about validity.

Three discourses of validity were identified. Validity as a *test characteristic* is underpinned by the notion that validity is an intrinsic property of a tool and could, therefore, be seen as content and context independent. Validity as an argument-based *evidentiary-chain* emphasizes the importance of supporting the interpretation of assessment results with ongoing analysis such that validity does *not* belong to the tool/instrument itself. The emphasis is on process-based validation (emphasizing the journey instead of the goal). Validity as a *social imperative* foregrounds the consequences of assessment at the individual and societal levels, be they positive or negative. The existence of different discourses may explain – in part – results observed in recent systematic reviews that highlighted discrepancies and tensions between recommendations for practice and the validation practices that are actually adopted and reported. Some of these practices, despite

contravening accepted validation ‘guidelines’, may nevertheless respond to different and somewhat unarticulated needs within health professional education.

TABLE 1: Summary of the three discourses

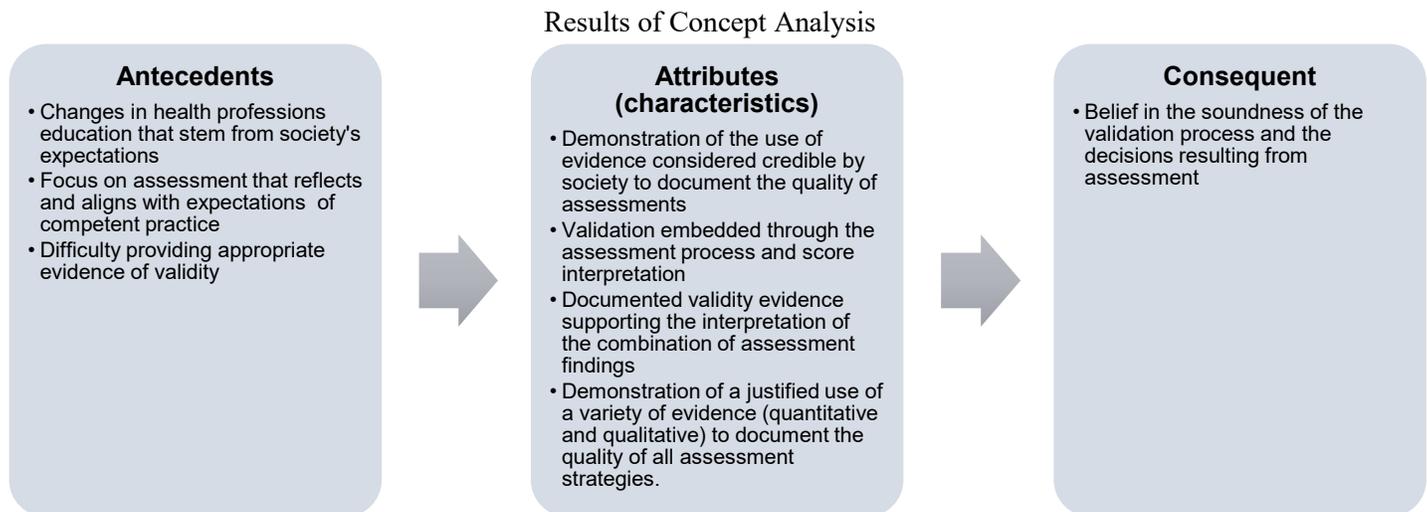
	Validity as a Test characteristic	Validity as an argument-based evidentiary-chain	Validity as a Social imperative
<i>Definition</i>	The degree to which the test actually measures what it purports to measure.	The evidences presented to support or refute the meaning or interpretation assigned to assessment results.	A bird’s eye view of assessment that foregrounds broader individual and societal issues
<i>Characteristics</i>	Validity is a goal or a gold seal of approval.	Validity is a journey on which one embarks to provide evidence supporting the interpretation of scores.	Validity and validation are matters of social accountability.
<i>Validity is viewed as...</i>	Static	Fluid	Built-in
<i>Focus of evidence is on...</i>	Individual tools can be considered valid, and the validity can generalize to the tool format (« MCQs are valid »)	Defensible interpretation of scores	Individual and societal impact of assessment
<i>Things made possible</i>	The quest for the holy grail of assessment; one tool that is more valid than the others.	Validation approaches and standards	Holistic and <i>a priori</i> consideration for societal impact of assessment
<i>Validation occurs...</i>	A posteriori (mainly)		A priori (mainly)
<i>Validation data focused on...</i>	Psychometric	Mostly psychometric	Mostly expert judgment

Context:

The adoption of the competency-based education (CBE) in the health professions education has been the catalyst for several changes in assessment, such as the use of assessment that mimics professional practice (authentic assessment) or the purposeful combination of different assessment strategies, contexts, and measurement times (programmatic assessment). Since assessment scores can have far-reaching consequences on future healthcare professionals and on society, it is essential to measure the quality (validity) of assessment strategies (Van der Vleuten & Schuwirth, 2005). However, there is a gap between the validation practices currently available to us and the current assessment approaches (Van der Vleuten & Schuwirth, 2005). This gap prompted reflections within the medical education literature, which were summarized in a recent discourse analysis (St-Onge et al., 2017). One of the results of this discourse analysis is a new conceptualization of validity: **validity as a social imperative**. Here, we explore this concept in order to describe it in the context of assessment in health professions education (HPE).

Concept analysis:

The first phase of this research was to explore validity as a social imperative using a concept analysis, according to the Rodgers' method (1989). The concept analysis is a useful method for identifying, clarifying, and fine-tuning an unexplored concept (Rodgers, 1989; Rodgers & Knafl, 2000). Roger's framework focuses on the reliance on literature to describe the antecedents, attributes, and consequents surrounding a concept. More concretely, *attributes* are the characteristics that define the concept (Rodgers & Knafl, 2000). *Antecedents* are what precedes the concept and the *consequences* are what happens as a direct result of the concept (Rodgers & Knafl, 2000).



*The attributes are described on next page

Now, it's your turn!

Since our current understanding of this conceptualization is mainly based on an in-depth analysis of the literature, your participation in this interview will allow us to explore the acceptability of the characteristics of the concept of **validity as a social imperative**.

Attributes (characteristics)	Description
<p>Demonstration of the use of evidence considered credible by society to document the quality of assessments</p>	<p>The various professional bodies (teaching institutions and professional orders) must be able to document with certainty—for society—the decisions made regarding the learners’ academic pathways and their level of competency for starting a professional career independently and competently.</p> <p>For example: A university is accountable to society for the decisions made based on the assessment of learners.</p> <p><i>“When students graduate from a university, the degree indicates to society that the graduates have a certain level of skill and expertise (Boley & Whitney, 2003, p.199).”</i></p>
<p>Validation embedded through the assessment process and score interpretation</p>	<p>When constructing an assessment program, elements which compose it must be chosen purposefully (Van der Vleuten et al., 2012). We should carefully consider how validity can be ‘built-in’ to the assessment process during its the development (Van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010). This consideration for validity throughout the assessment development process increases the credibility, defensibility and accuracy of the score interpretation (Berendonk, Stalmeijer, & Schuwirth, 2013).</p> <p><i>Ebel (Ebel, 1983) also argued that validity can be a ‘built-in’ feature of an assessment method. We take the view that all assessment at the three bottom layers of Miller’s pyramid can be controlled and optimised: materials can be scrutinised, stakeholders prepared, administration procedures standardised, psychometric procedures put in place, etc (Van der Vleuten, Schuwirth, Scheele, Driessen, & Hodges, 2010, p.707).</i></p> <p>Another element to be considered during the development of the assessment program is the consequences of the assessment process and subsequent score interpretation. The person responsible for the assessments must anticipate the potential consequences and implement measures or strategies to minimize them. The consequences measured should not be solely limited to the impacts of the construct, but rather all possible consequences.</p> <p><i>“(…) the measurement or scoring procedure (e.g., irrelevant, unreliable, or omitted test items); the specific interpretation (e.g., an inappropriate pass/fail cut point); the attribute being measured (i.e., the wrong construct); or the response (e.g., the actions that follow the decision) (Cook & Lineberry, 2016, p.790).”</i></p>
<p>Documented validity evidence supporting the interpretation of the combination of assessment findings</p>	<p>The interpretation of assessment scores must be done from the perspective of a “whole” (the assessment program in its entirety) that is greater than the sum of its parts.</p> <p><i>“The central key is that the programme of assessment is set up to allow the whole picture of a student’s competence to be obtained by a careful selection of assessment methods, formulation of rules and regulations and design of organisational systems (Schuwirth & Van der Vleuten, 2011, p.479).”</i></p>
<p>Demonstration of a justified use of a variety of evidence (quantitative and qualitative) to document the quality of all assessment strategies.</p>	<p>Since traditional quantitative analysis (e.g., Cronbach’s alpha, psychometric analysis, etc.) are often lacking applicability for demonstrating the quality of a set of assessment strategies (i.e., assessment program), the combination of quantitative and qualitative evidences appears to be a solution to be considered.</p> <p><i>For qualitative assessments, the synthesis of individual pieces of qualitative data to form an insightful, accurate and defensible interpretation is analogous to quantitative generalisation. Whereas we treat inter-rater variability as error for most numeric scores, in qualitative assessments we view observer variability as representing potentially valuable insights into performance (i.e. different perspectives). The method for selecting and synthesising data from different sources (triangulation) and deciding when to stop (saturation) will inform the Generalisation inference for qualitative data. (Cook, Brydges, Ginsburg, & Hatala, 2015, p.567).</i></p>