Université de Sherbrooke

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci

Par
Élise Vachon Lachiver
Programme recherche en sciences de la santé
Cheminement en pédagogie des sciences de la santé

Mémoire présenté à la Faculté de médecine et des sciences de la santé
en vue de l'obtention du grade de maitre ès sciences (M. Sc.)
en sciences de la santé

Sherbrooke, Québec, Canada
Mars 2017

Membres du jury d'évaluation
Pre Christina St-Onge, PhD, directrice, Programme recherche en sciences de la santé
Pre Cécile Trochet, M.D., M. Sc., membre interne, Programme recherche en sciences de la santé
Pr Pascal Ndinga, PhD, membre externe, Département d'éducation et pédagogie, Faculté des sciences de l'éducation, UQAM

RÉSUMÉ

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci

Par
Élise Vachon Lachiver
Programmes recherche en sciences de la santé

Mémoire présenté à la Faculté de médecine et des sciences de la santé en vue de l'obtention du diplôme de maitre ès sciences (M.Sc.) en sciences de la santé, Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, Québec, Canada, J1H 5N4

Introduction : L'évaluation des apprentissages a une importance indéniable en pédagogie des sciences de la santé et l'utilisation d'examens écrits est omniprésente, notamment en médecine. Toutefois, l'élaboration d'évaluations de haute qualité est souvent un défi pour les rédacteurs et les programmes. Il est donc important d'identifier des stratégies qui pourraient faciliter le travail des rédacteurs tout en ayant un impact positif sur la qualité des questions écrites, notamment l'utilisation de consignes lors de la rédaction.

Méthodologie : Nous avons élaboré une liste de consignes de rédaction de questions d'examens par un processus itératif et rigoureux. 14 consignes discriminantes ont été retenues. À l'aide d'une étude de cas multiples, nous avons exploré l'utilisation de ces consignes de rédaction de questions auprès de quatre rédacteurs pour des examens d'un programme préclinique de médecine. Un questionnaire et une entrevue individuelle semi-dirigée ont été effectués pré et post utilisation des consignes. Les unités d'analyses et de comparaison étaient : la qualité et les propriétés psychométriques des questions pré et post utilisation des consignes, l'acceptabilité et l'utilisation de celles-ci.

Résultats : Pour tous les participants, les questions rédigées respectaient majoritairement les consignes de rédaction. Quoique non statistiquement significative, nous avons observé une légère augmentation du coefficient de discrimination chez les rédacteurs (2 sur 4) qui utilisaient les consignes de manière active. Nos résultats suggèrent que l'appropriation que font les rédacteurs des consignes semble influencer leur perception de l'utilité et de l'acceptabilité des consignes.

Conclusion : Les résultats obtenus semblent suggérer que l'utilisation d'une liste circonscrite de consignes de rédaction offre une bonne stratégie à adopter pour améliorer la qualité des questions. Les rédacteurs de questions qui sont mieux outillés et qui semblent adopter pleinement l'utilisation des consignes de rédaction y voient les impacts positifs sur la qualité de leurs questions.

Mots clés : Évaluation, Examens écrits, Rédacteurs de questions, Propriétés psychométriques, Étudiants en médecine

SUMMARY

The use of guidelines for writing exam questions in medicine and the effect on the psychometric properties of the exams

By
Élise Vachon Lachiver
Health sciences Program

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

Introduction : The assessment of learners has an undeniable importance in health professions education and the use of written exams is omnipresent, notably in medicine. However, developing high-quality assessment is often a challenge for administrators and exam developers. It is therefore important to identify strategies that could facilitate their work while having a positive impact on the quality of the written exam questions, including the use of guidelines when developing exams.

Methodology : We created a list of item-writing guidelines through a rigorous and systematic process. We identified 14 discriminant guidelines. Using a multiple case study, we explored the use of those questions writing guidelines with four exam developers at the preclinical phase of an undergraduate medicine program. A questionnaire was completed and a semi-directed individual interview was carried out before and after the use of guidelines. The units of analysis and comparison were: the quality and psychometric properties of questions pre- and post-use of the guidelines, the acceptability and the use of these guidelines.

Results : For all the participants, the written questions respected mostly the instructions of drafting. Although not statistically significant, we observed a slight increase in the coefficient of discrimination among editors who actively used the instructions. Our results suggest that participants' appropriation of the item-writing guidelines seemed to influence their perception of the usefulness and acceptability of instructions.

Conclusion : The results seem to suggest that the use of a narrow list of writting guidelines provides a good strategy for improving the quality of the questions. Exam developers that are better equipped and seem to fully adopt the use of drafting guidelines see the positive impacts on the quality of their questions.

Keywords : Evaluation, Written exams, Exam developers, Psychometric properties, Medical student

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

Figures de l'article 1

Figures de l'article 2

LISTE DES ABRÉVIATIONS

AE	*Assessment Engineering*
HPE	*Health professions education*
MCQ	*Multiple choice question*
MD	Médecine
QCM	Question à choix multiples
QROC	Question à réponse ouverte courte
SAQ	*Short-answer question*
UdeS	Université de Sherbrooke

À Maman, Papa, mes sœurs et Simon

REMERCIEMENTS

Le premier remerciement se doit d'être pour ma directrice, Pre Christina St-Onge, celle qui m'a permis de faire mes premiers pas dans le monde fleurissant de la pédagogie médicale. Grâce à son soutien, son écoute, ses conseils et son expérience, j'ai pu réaliser ce projet de recherche à travers les hauts et les bas que cela comporte. Je lui suis également reconnaissante pour les multiples opportunités qu'elle m'a apportées et pour tout ce qu'elle m'a apprise. Je me considère choyée d'avoir mis à terme ce premier projet avec toi. Un énorme merci pour tout.

Je remercie également toute l'équipe de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des Médecins de l'Université de Sherbrooke avec qui j'ai la chance de travailler et de partager des biscuits! Merci Linda de ton aide et tes commentaires toujours constructifs. Kathleen, merci notamment pour ton soutien avec la portion qualitative de ce projet. Merci aussi à Marianne, Martine et Daniel pour vos commentaires et recommandations. Votre implication et votre amour de la pédagogie médicale ont su résonner en moi.

Je tiens à remercier le programme de médecine de l'Université de Sherbrooke pour la collaboration dans ce projet de recherche.

Dans ce sens, je veux également souligner la participation des rédacteurs de questions qui se sont prêtés au jeu; sans eux, vous n'auriez pas grand-chose à lire.

Merci à mes collègues, Mélanie, Élisabeth et Hassiba pour les discussions et les échanges, ils m'ont permis de relativiser les moments plus difficiles et de partager les bons coups. Après tout, on est tous dans le même bateau traversant cet océan que sont les études supérieures!

Finalement, je remercie ma famille pour ses encouragements et son support tout au long de ma maîtrise. Vous avez su, chacun à votre manière et malgré les kilomètres qui nous séparent, me motiver à me rendre jusqu'ici. Merci Maman pour ton amour, ta compréhension et tes encouragements. Merci Papa, peu de mots suffisent à qui sait comprendre. Merci à mes chères sœurs Émilie, Édith et Élodie, vous m'apportez énormément chacune à votre façon. Simon, merci pour ton soutien, ta patience dans les moments plus stressants et ta volonté à faire en sorte que je me dépasse de jour en jour. Encore une fois, je vous dis merci. Vous êtes en partie responsables de ce succès.

INTRODUCTION

L'évaluation des apprenants est définie comme un processus qui consiste à porter un jugement sur les apprentissages, à partir de données recueillies, analysées et interprétées, en vue de décisions pédagogiques et administratives (Jouquan, 2002; Scallon, 2004). L'évaluation nous permet de justifier les décisions et les actions pédagogiques qui régulent les apprentissages des étudiants à différents moments dans le parcours scolaire (Brady, 2005; van der Vleuten, 2000). L'évaluation des apprenants remplit donc une fonction capitale. Cela est tout aussi vrai dans les programmes de pédagogie des sciences de la santé et de médecine (Case et Swanson, 1998). L'évaluation des apprenants est au cœur des apprentissages tant au plan préclinique qu'à l'externat ou à la résidence des programmes de médecine. C'est dans le contexte préclinique de ces programmes que nous allons principalement nous attarder.

L'évaluation des apprenants dans les programmes de médecine a des enjeux de taille à respecter. D'une part, les évaluations doivent vérifier que les étudiants ont acquis les connaissances qui leur ont été enseignées et qui respectent les objectifs du programme (Case et Swanson, 1998). D'autre part, les évaluations sont utilisées comme point de contrôle auprès des étudiants en leur donnant accès aux étapes futures du programme. Allant jusqu'à leur permettre l'accès à la profession souhaitée, soit celle de médecin (Epstein et Hundert, 2002). Les évaluations peuvent d'endosser ce rôle quand elles parviennent à discriminer les étudiants qui possèdent les connaissances adéquates et les habiletés nécessaires pour se retrouver en milieu clinique (Brady, 2005; Linn et Gronlund, 2000; van der Vleuten, 2000).

Il est connu dans le milieu de la pédagogie médicale que la rédaction de questions d'examen est une tâche difficile (Bush, 2006; Tarrant *et al.*, 2009). Cela représente un défi de taille pour les rédacteurs, certains diront même qu'il s'agit autant d'un art que d'une science (Downing, 2003; Haladyna *et al.*, 2002). En effet, la rédaction de questions d'examens écrits demande beaucoup de temps, d'efforts, de ressources et de connaissances

de la part des rédacteurs de questions (Caldwell et Pate, 2013; Downing, 2005). Vu la difficulté de la tâche, il arrive parfois que des questions problématiques au niveau statistique (docimologie) se retrouvent dans les évaluations. La littérature scientifique, contrairement à ce qu'on pourrait penser, n'aide pas les rédacteurs dans leur rédaction de questions. Effectivement puisqu'il existe un très grand nombre de consignes de rédaction pour les questions à choix multiples et les questions à réponse ouverte courte, il peut ainsi être difficile pour eux de savoir laquelle de ces consignes les rédacteurs doivent utiliser. La rédaction de questions d'examens écrits demeure donc quelque chose d'assez complexe à faire pour les rédacteurs.

De plus, il existe peu de données empiriques qui nous indiquent l'impact du respect des consignes sur les propriétés psychométriques des questions (Downing, 2005; Haladyna et Downing, 1989; Prihoda *et al.*, 2006). Autrement dit, on ne connait pas l'impact de l'utilisation des consignes de rédaction au niveau de la qualité de celles-ci. On se demande donc si l'utilisation de consignes pendant le processus de rédaction pourrait d'une part faciliter le travail des rédacteurs tout en ayant aussi un impact positif sur la qualité des questions écrites. Cette stratégie pourrait venir augmenter la qualité des évaluations sans atteinte à la validité en étant utilisée a priori.

Ce mémoire porte sur une étude réalisée auprès de rédacteurs de questions du programme de médecine de l'Université de Sherbrooke qui sont responsables de rédiger des questions d'examens écrits dans la phase préclinique du programme. Tout d'abord, la problématique étudiée expose les enjeux reliés à la rédaction d'évaluation de qualité dans le contexte de la pédagogie des sciences de la santé, notamment en médecine. Par la suite, la recension des écrits est présentée, l'existence de consignes de rédaction de questions ainsi que de leur impact sur la qualité des évaluations sont décrits. La première partie du mémoire se termine par la description des objectifs de l'étude. Les sections subséquentes présentent deux articles qui rapportent la méthodologie de l'étude, les résultats obtenus, et leur interprétation. Puis, la présentation d'une discussion qui intègre les résultats exposés dans les articles, aborde les forces et limites de l'étude, et propose des perspectives de recherches, avant de conclure ce mémoire.

Premier chapitre : PROBLÉMATIQUE

L'évaluation des apprentissages a été définie comme un processus incluant un jugement basé sur des données recueillies qui ont été analysées et interprétées dans le but de rendre une décision pédagogique et administrative (Jouquan, 2002; Scallon, 2004). Elle permet de justifier les décisions et les actions qui régulent les apprentissages de l'élève, dans le quotidien comme à des moments plus stratégiques. Ainsi, l'évaluation des apprentissages remplit une fonction pédagogique capitale (Gouvernement du Québec, Ministère de l'Éducation, 2003). L'évaluation des apprentissages et des compétences est au cœur de la pédagogie des sciences de la santé, tant au plan préclinique qu'à l'externat ou à la résidence dans les programmes de médecine (Case et Swanson, 1998).

Les évaluations des apprentissages dans les programmes de médecine ont un enjeu de taille à considérer. En effet, elles sont utilisées notamment pour vérifier que les étudiants ont acquis les connaissances leur étant enseignées et qui respectent les objectifs du programme (Case et Swanson, 1998). Les évaluations sont notamment utilisées pour assurer un premier point de contrôle auprès des étudiants, leur donnant accès aux étapes subséquentes dans le cursus. Par exemple, l'évaluation des apprentissages en contexte préclinique des programmes de médecine doit permettre d'identifier les étudiants qui possèdent les connaissances adéquates et les habiletés nécessaires pour accéder à la prochaine étape soit les stages en milieu clinique (Brady, 2005; Epstein et Hundert, 2002; Linn et Gronlund, 2000; van der Vleuten, 2000). L'accès à la profession souhaitée est un enjeu majeur et une conséquence directe des évaluations. Pour prendre de bonnes décisions et minimiser les conséquences négatives de l'évaluation, les programmes s'appuient donc sur des évaluations de qualité. Ces dernières permettraient de discriminer les étudiants plus faibles qui pourraient se retrouver en situation d'échec (Jouquan, 2002). Pour répondre à ces différents objectifs de la formation médicale, le développement d'examens doit être fait de manière rigoureuse et appropriée (Case et Swanson, 1998) et rencontrer les plus hauts standards de qualité. Ainsi, les évaluations effectuées au niveau préclinique du programme

de médecine sont utilisées pour faire le suivi auprès des étudiants et repérer ceux qui présentent des difficultés.

Il est connu dans le milieu de la pédagogie médicale que l'élaboration et la rédaction de questions d'examen sont ardues (Bush, 2006; Haladyna *et al*., 2002; Tarrant *et al.*, 2009). L'élaboration de questions d'examens écrits est souvent décrite comme étant autant un art qu'une science (Downing, 2003; Rodriguez, 2005). Elle représente un défi de taille. En effet, la rédaction de questions d'examens écrits demande beaucoup de temps, d'efforts, de ressources et de connaissances de la part des rédacteurs de questions (Caldwell et Pate, 2013). Une option qui s'offre aux rédacteurs quant à la rédaction de questions de qualité est de se tourner vers les écrits scientifiques où l'on retrouve une panoplie de consignes de rédaction (Case et Swanson, 1998; Considine *et al.*, 2005; Downing, 2005; Haladyna et Downing, 1989; Moreno *et al.*, 2006; Prihoda *et al.*, 2006). Toutefois, avec plus de 150 consignes répertoriées, il peut être difficile pour les rédacteurs de s'y retrouver faute d'expertise et de ressources. De plus, il existe peu de données empiriques quant à ces consignes, c'est-à-dire qu'il y a peu de données nous indiquant si le respect ou non-respect, de ces consignes a un impact sur les propriétés psychométriques des questions d'examens écrits. Il est ainsi encore plus difficile pour les rédacteurs de naviguer à travers cette panoplie de consignes recensées dans les écrits scientifiques.

Malheureusement, il arrive que des questions d'examens soient mal rédigées (problèmes de contenu, ambigüités, faibles propriétés psychométriques, etc.) et se retrouvent dans l'évaluation. Le monitorage de la qualité des évaluations est ainsi mis en place pour s'assurer que les examens respectent un certain niveau de qualité. Pour ce faire, différentes stratégies de monitorage existent notamment l'analyse d'items. Cette stratégie utilisée suite à l'administration de l'examen, permet d'identifier les questions mal construites ou problématiques afin de les éliminer lorsqu'il y a lieu (Downing, 2002; Haladyna *et al.*, 2002). L'analyse d'items est ainsi largement utilisée dans les divers programmes de sciences de la santé.

L'utilisation de l'analyse d'items comme stratégie de monitorage (modification *a posteriori*) entraine néanmoins des conséquences négatives sur la validité de l'interprétation des résultats de l'évaluation. La validité selon Bertrand et Blais (2004) « consiste en un jugement basé sur des preuves empiriques et sur une argumentation de nature théorique qui vise à justifier l'interprétation des scores obtenus à la suite de l'administration d'un test dans un contexte donné ». Lors de l'élaboration d'un examen, les rédacteurs doivent habituellement élaborer des questions représentatives d'un certain contenu. L'élimination de questions d'examens *a posteriori* diminue la représentativité du contenu devant initialement être mesuré. En effet, le contenu modifié n'est parfois plus complètement représentatif du contenu devant être évalué. Cette modification au plan de la représentativité du contenu peut s'avérer problématique, car de l'information considérée initialement comme importante manque dans la version finale de l'examen. D'autant plus qu'il existe peu de stratégies de monitorage de la qualité des évaluations faites en aval de la correction de questions.

L'utilisation de consignes lors du processus de rédaction pourrait faciliter le travail des rédacteurs tout en ayant un impact positif sur la qualité des questions écrites. Une des stratégies qui mérite d'être explorée est celle d'outiller les rédacteurs de questions au moment même où ceux-ci écrivent leurs nouvelles questions. Cela signifie d'utiliser *a priori* les consignes de rédactions de questions d'examens écrits et ainsi éviter de diminuer la validité de l'évaluation. Il faut donc préalablement identifier les consignes de rédaction de questions qui permettent de discriminer les questions d'examens écrits de haute et de basse qualité. Ainsi, l'objectif général de cette recherche est d'identifier les consignes de rédaction discriminantes et de les utiliser comme outils pour les rédacteurs lors de la rédaction de questions d'examens écrits.

Deuxième chapitre : RECENSION DES ÉCRITS

L'évaluation des apprentissages est certainement un élément important de tout programme de formation. L'évaluation des apprentissages des étudiants est même considérée aujourd'hui comme l'un des éléments essentiels du processus enseignement-apprentissage. Les cadres théoriques qui inspirent ce concept rendent compte des divers courants de la pédagogie. La locution anglophone « *assessment drives learning* » en est d'ailleurs un bon exemple (Wood, 2009, p.5). Quant aux pratiques évaluatives, elles sont elles aussi fortement influencées par un ou plusieurs courants conceptuels qui en constituent les fondements et qui permettent d'analyser à la fois les forces et les limites de la pratique évaluative utilisée. En plus d'être très diversifiées quant aux formats qu'elles adoptent et aux outils qu'elles utilisent, les pratiques évaluatives répondent aux différents objectifs des programmes (Jouquan, 2002).

2.1. Les différents enjeux de l'évaluation

L'essentiel d'une évaluation selon Scriven (1967, traduit et cité dans Auger, 2000, p.27) est « l'intention qui amène une décision à entreprendre une action à la suite d'un ensemble d'informations recueillies ». Pour ce faire, les programmes utilisent à travers le cursus scolaire et en fonction du contexte, des évaluations formatives, sommatives ou encore des évaluations de certification. Scallon (2000) définit l'évaluation formative de la manière suivante :

> *[…] Processus d'évaluation continue ayant pour objectif d'assurer la progression des individus engagés dans une démarche d'apprentissage ou de formation, selon deux voies possibles : soit par des modifications de la situation ou du contexte pédagogique, soit en offrant à chaque individu l'aide dont il a besoin pour progresser, et ce, dans chacun des cas, pour apporter, s'il y a lieu, des améliorations ou des correctifs appropriés.*
>
> (p.21)

Pour ce qui est de l'évaluation sommative, Scallon (2000, 2004) mentionne qu'en plus de l'idée de somme ou d'addition contenue dans l'expression, elle fait également référence à la

reconnaissance d'un niveau de développement au terme d'une période donnée. Le National Research Counsil (NRC) proposait en 2001 des définitions semblables qui différenciaient l'évaluation qui vise à soutenir l'apprentissage (formative assessment) de l'évaluation qui vise à déterminer un niveau de développement (summative assessment) (Tardif, 2006). Les évaluations formatives et certaines évaluations sommatives, dépendamment du contexte où elles sont administrées aux étudiants, font partie des évaluations à faibles enjeux (*low-stakes assessment*). Cela veut dire que la réussite ou l'échec de l'évaluation à de plus faibles répercussions sur le parcours académique de l'étudiant. Il existe des situations où des évaluations sommatives possèdent des enjeux élevés (*high-stakes assessment*), notamment au moment de la transition entre le préclinique et l'externat. Les évaluations de certification ainsi que l'admission des étudiants dans les programmes de médecine font également partie des évaluations à enjeux élevés (Scallon, 2004). Ces évaluations se distinguent par le fait que les conséquences qui y sont rattachées sont plus importantes. En effet, le but de l'évaluation de certification n'est pas de déterminer le niveau d'acquisition des apprentissages des étudiants, mais bien de sanctionner le niveau d'apprentissage pouvant, par exemple, donner accès à la profession souhaitée. Ainsi, selon le contexte dans lequel se déroule l'évaluation les enjeux qui s'y rattachent varient allant de faibles à élevés.

2.2. Standards de qualité des évaluations

Nous utilisons des standards pour favoriser ou promouvoir la qualité des résultats obtenus, et pour s'assurer qu'ils répondent aux besoins de chacune des parties prenantes. En pédagogie des sciences de la santé, un certain nombre d'intervenants différents sont impliqués ou affectés par les évaluations et leurs résultats. Les parties prenantes comprennent les patients, le grand public, les employeurs de la santé, les organismes professionnels et règlementaires, les universités, les écoles de médecine, les organismes de formation, les enseignants individuels et, enfin et tout aussi important, les candidats eux-mêmes (Amin *et al*,. 2006). Chacune de ces parties prenantes fait donc un usage différent d'une même évaluation et peut alors avoir différentes priorités quand vient le temps de prioriser certains standards. Ainsi, aucun ensemble de standards de qualité d'une bonne évaluation ne s'applique bien à toutes les situations. En fait, les mêmes standards devraient avoir une importance différente selon l'objectif et le contexte de l'évaluation. Norcini et

collaborateurs (2011), par un consensus de groupe, ont établi sept standards de qualité pour une bonne évaluation. Il s'agit entre autres de la validité ou de la cohérence, de la reproductibilité ou constance, de l'équivalence, de la faisabilité, de l'effet éducatif, de l'effet catalytique ainsi que de l'acceptabilité de l'évaluation. Comme mentionné plus haut, il est important de choisir les standards de qualité qui doivent être respectés selon le type d'évaluation et le contexte dans lequel elle se déroule.

2.3. Évaluation de la qualité des questions et stratégies employées

Un contrôle méthodique et serré est fait sur les examens en contexte de formation médicale pour s'assurer de la qualité des évaluations. Cette pratique est même régie par les organismes responsables de l'agrément des programmes de médecine (CACMS Standards and Elements, 2015). Qui plus est, différentes associations aux États-Unis, au Canada et au Québec ont élaboré des normes et des standards de qualité à atteindre lors de l'élaboration et de l'utilisation de procédés d'évaluation (outils et instruments) des apprentissages (American Psychological Association, American Educational Research Association, National Council on Measurement in Education, Comité consultatif mixte pour le Canada et la politique d'évaluation du ministère de l'Éducation du Québec). Ces divers documents visent tous à répondre à une préoccupation commune, celle d'assurer aux procédés et aux outils de mesure et d'évaluation les qualités nécessaires afin de respecter les règles élémentaires d'équité, de justice et d'égalité (Scallon, 2004). Les principes de justice et d'équité apparaissent comme deux principes importants dans les politiques d'évaluation. Des efforts ont été fournis en psychométrie classique pour standardiser les conditions d'évaluation dans le but de s'assurer qu'aucun étudiant ne bénéficie d'une situation au détriment d'un autre. Il s'agit de sélectionner les contenus d'apprentissage, les stratégies, les méthodes et toute autre intervention éducative de manière à tenir compte des besoins de l'ensemble des étudiants. Ce souci est d'autant plus important dans les évaluations à enjeux élevées (Scallon, 2004).

Plus concrètement, il existe plusieurs stratégies de monitorage pour juger la qualité et l'atteinte des standards pour chacune des évaluations. Une de ces stratégies mises en place notamment dans les programmes de médecine est l'analyse d'items. Les examens en milieu

scolaire ou universitaire sont rarement mis à l'épreuve avant la passation en classe par les étudiants. Ceci rend l'analyse d'items d'autant plus essentielle puisque c'est le moyen dont dispose le personnel enseignant pour modérer les résultats d'examen. On peut concevoir l'analyse d'item comme un processus de dépistage où les « mauvais » items sont retirés du score total. Les bonnes questions sont, quant à elles, conservées, car elles nous permettent de répondre à notre objectif d'évaluation. Pour nous aider à faire cette sélection, quatre critères peuvent être utilisés; l'indice de difficulté et l'indice de discrimination des questions ainsi que la fidélité et la validité des évaluations. La fidélité et la validité font référence à l'ensemble des questions contenues dans les évaluations alors que pour les deux premiers critères (indices de difficulté et de discrimination), ils sont propres à chacune des questions. Les techniques d'analyse d'items sont nombreuses et variées et doivent être choisies en fonction du type d'évaluation (formative ou sommative, par exemple). Cela nous fournira des informations sur le rôle de chacune des questions sur le score total de l'évaluation (Laveault et Grégoire, 2014).

Une première étape pour juger de la qualité des évaluations est de regarder les propriétés psychométriques des questions, soit les indices de discrimination et les indices de difficultés des questions. L'indice de discrimination représente la capacité d'une question à discriminer les candidats de faibles et ceux de fortes habiletés. Cette dernière variable étant perçue comme étant une mesure des compétences générales du candidat. L'indice de discrimination peut se situer dans un intervalle compris entre -1 et +1. Pour qu'une question puisse permettre de distinguer efficacement les candidats faibles des candidats forts, le résultat de ces derniers doit être supérieur. Dans ce cas, l'indice de discrimination sera généralement compris entre 0,1 et 0,4. En règle générale, les questions qui permettent de départager efficacement les candidats forts des candidats faibles ont un indice de discrimination positif et supérieur à 0,2 (Conseil médical du Canada, 2010). Ces questions rehaussent l'efficacité générale de l'examen pour ce qui est de départager les bons candidats des candidats plus faibles. Ebel et Frisbie (1986) proposent des repères pour interpréter l'indice de discrimination (Tableau 1).

Tableau 1. Barèmes proposés pour l'interprétation de l'indice de discrimination

Indice de discrimination	Interprétation
> 0.40	Item qui discrimine très bien
0.30 - 0.39	Item qui discrimine bien
0.20 - 0.29	Item qui discrimine peu
0.10 - 0.19	Item-limite, à améliorer
< 0.10	Item sans utilité réelle pour l'examen

Ces indications sont souvent utilisées lors d'examen à résultats dichotomiques ou encore lors des évaluations à enjeux élevés, tel que les examens de certification.

Pour une question à réponse dichotomique, l'indice de difficulté représente la proportion d'individus qui réussissent la question. Plus l'indice augmente, plus la question est considérée comme facile. L'indice de difficulté varie de 0 à 1,0 (Crocker et Algina, 1986). Ainsi, une question dichotomique avec un indice de difficulté de 0,75 indique que 75 % des individus ont répondu correctement à la question. Dans le cas d'une question cotée sur une échelle de plusieurs points, « l'indice de difficulté nous est indiqué par la moyenne des cotes accordées à cet item pour l'ensemble des sujets » (Laveault et Grégoire, 2014, p.204). Une formule nous permet ainsi d'obtenir une valeur variant entre 0 et 1 comme si la question était dichotomique. Il est également important de considérer deux facteurs pouvant influencer notre interprétation de l'indice de difficulté soit, le nombre de réponses omises (étudiants n'ayant pas répondu à la question) et la probabilité de réussir la question due au hasard. Lorsque l'indice de difficulté est calculé pour une QCM, on doit alors tenir compte de cette probabilité. Pour ce faire, une formule de correction de l'indice de difficulté pour le hasard est utilisée chaque fois que l'on considère que les leurres ont une chance égale d'être choisis (Laveault et Grégoire, 2014).

L'analyse d'item vise à éliminer les questions problématiques (questions mal formulées, ambigües, problème de contenu, etc.) des évaluations dans le but d'en assurer la qualité. Ainsi, l'utilisation de l'analyse d'items comme stratégie de monitorage peut entrainer des conséquences négatives sur la validité de l'interprétation des résultats de l'évaluation. En se basant sur la définition de la validité de Bertrand et Blais (2004) mentionnée plus haut, on comprend donc que lorsqu'il y a des questions qui sont éliminées d'une évaluation, la

représentativité du contenu global de l'évaluation est diminuée. Nous ne sommes plus en train d'évaluer le contenu qui devait initialement être mesuré. Les données concernant les connaissances des étudiants sur le contenu qui a été retiré des évaluations sont donc inconnues des évaluateurs.

Ainsi, malgré les stratégies d'évaluation des questions dans le but de satisfaire les standards de qualité, la rédaction de questions d'examens écrits demeure une tâche difficile pour les rédacteurs. Certains auteurs ont donc tenté de faciliter ou baliser la tâche de rédaction en formulant diverses consignes de rédaction. Une recension de la littérature scientifique nous a permis de dresser un portrait actuel des connaissances de la rédaction de questions d'examens écrits en s'intéressant principalement aux consignes de rédaction.

2.4. Consignes de rédaction de questions

Les résultats de cette recension nous permettent de tirer trois constats. Le premier est qu'il existe un grand nombre de consignes de rédaction, plus de 150 consignes ont été répertoriées dans la littérature scientifique (St-Onge *et al.*, 2013). Deuxième constat, beaucoup d'études portent sur l'élaboration ou l'évaluation des consignes de rédaction *a posteriori*, mais aucune d'entre elles ne s'intéresse à l'utilisation *a priori* des consignes de rédaction de questions. Troisième constat, ce ne sont pas toutes les consignes qui font l'objet d'une analyse de leur impact sur la qualité des questions. Parmi les articles retenus, seulement quatre consignes de rédaction ont été analysées par rapport à leur effet sur la qualité des questions. Il vous sera présenté en détail un peu plus loin chacune des quatre consignes pour lesquelles des analyses de l'impact du respect ou non de leur utilisation ont été faites. D'abord, explorons chacun de ces trois constats qui ressortent plus en détail.

2.4.1. Premier constat : L'existence d'un grand nombre de consignes

La littérature scientifique regorge de consignes de rédaction de questions d'examens écrits. Certaines de ces consignes font l'objet d'un fort consensus de la part des auteurs tandis que d'autres consignes sont moins connues. Il existe également des consignes de rédaction (*guidelines*) pour lesquelles des désaccords existent parmi les auteurs (Haladyna et Downing, 1989). Cela s'explique peut-être en partie par le fait que les auteurs qui

s'appuient sur des expériences personnelles ainsi que sur des recherches empiriques limitées ne justifient que rarement les consignes qu'ils présentent, que cette justification soit logique ou empirique (Haladyna *et al.*, 2002). L'étude de Haladyna et Downing (1989) avait comme objectif de faire une taxonomie des consignes de rédaction de question à choix multiples. Chacune des 43 consignes retenues fait l'objet d'un consensus de la part des auteurs. L'atteinte de ce consensus amène dans un sens un poids à la pertinence de cette consigne puisque les auteurs possèdent un haut niveau d'expertise en évaluation. Cette étude a permis, grâce à sa méthodologie rigoureuse, de réduire le nombre de consignes de rédaction qui devraient être considérées ou utilisées par les rédacteurs de questions. Cependant lorsqu'on consulte la littérature scientifique sur le sujet, on s'aperçoit rapidement que d'autres auteurs continuent de publier leurs propres consignes de rédaction maison en vérifiant seulement *a posteriori* si les questions rédigées respectent ces mêmes consignes.

2.4.2. Deuxième constat : Utilisation majoritairement a posteriori des consignes

On rapporte deux types d'études portant sur les consignes de rédaction de questions; d'une part des études qui élaborent de nouvelles consignes de rédaction et d'autre part des études qui portent sur l'évaluation de consignes de rédaction. Ce deuxième type d'études vise à évaluer l'utilisation de consignes de rédaction sur la qualité des questions d'examens (Bush, 2006; Case et Swanson, 1998; Considine *et al.*, 2005; Downing, 2005; Downing, 2002; Haladyna et Downing, 1989; Moreno *et al.*, 2006; Prihoda *et al.*, 2006; Rodriguez, 2005; van der Vleuten, 2000). Cette évaluation des consignes se fait une fois que les questions écrites d'examens ont été répondues par les apprenants. Conséquemment, l'intérêt de l'utilisation des consignes de rédaction dans la littérature scientifique est de vérifier la qualité des questions *a posteriori*. Les études ne s'intéressent toutefois pas au potentiel de l'utilisation des consignes de rédaction de questions pendant la rédaction comme pouvant contribuer à la qualité psychométrique de l'examen. En effet, aucune étude ne s'intéresse à l'utilisation des consignes au moment de la rédaction, avant même que l'évaluation ait été administrée aux étudiants.

2.4.3. Troisième constat : Respect des consignes et l'effet sur la qualité des questions

Il apparait rapidement que ce ne sont pas toutes les consignes de rédaction qui font l'objet d'une analyse de leur impact sur la qualité des questions. Le non-respect ou la violation de consignes de rédaction a comme conséquences de formuler des questions imparfaites et cela entraine également des effets négatifs quant à la qualité de ces questions (Downing, 2005; Downing, 2002). Les paragraphes qui suivent démontrent plus en détail comment l'utilisation de consignes de rédaction peut influencer la qualité des questions. De cette recension des écrits, seulement quatre consignes de rédaction ont des données probantes.

2.4.3.1. Leurres mal construits

Une des consignes de rédaction qui a été évaluée parmi les études recensées concerne les leurres mal construits. Une question à choix multiple possède une bonne réponse ainsi qu'un certain nombre de leurres. Un leurre agit comme source de diversion et est souvent attirant pour les étudiants plus faibles. Autrement dit, un leurre est une option de réponse erronée.

La présence d'un ou plusieurs leurres mal construits entraine différents impacts sur la qualité des questions, mais également sur les évaluations de manière plus générale. Les indices de discrimination et de difficulté peuvent être affectés tout comme la reproductibilité ce qui a pour effet de modifier la note totale de l'évaluation (Downing, 2002). Le seuil de passage est également influencé par la présence de leurres mal construits dans l'évaluation. À partir d'un examen de première année de médecine portant sur des notions en science de base, Downing (2002) a observé qu'un tiers des questions (11 sur 33) comportait un leurre mal construit. L'indice de difficulté des questions avec des leurres mal construits a diminué de plus de la moitié d'un écart-type comparativement à une question bien formulée évaluant le même contenu (provenant d'une autre évaluation). Cette augmentation de la difficulté des questions a eu comme répercussion de faire échouer presque un quart d'étudiants de plus, malgré une diminution de deux points du seuil de passage.

Tarrant et collaborateurs (2009) ont voulu quantifier la proportion de leurres mal construits dans 514 questions à choix multiples de sept examens d'un programme de soins infirmiers. Les auteurs ont aussi regardé la difficulté et la discrimination rattachées aux questions avec des leurres mal construits. La proportion de questions comportant 0, 1, 2 ou 3 leurres mal construits était respectivement de : 13,8 %, 39,1 %, 34,8 % et 12,3 %. En ce qui a trait à la difficulté des questions, les résultats de cette étude vont à l'encontre des résultats de Downing (2002); les questions avec 2 ou 3 leurres mal construits étaient significativement moins difficiles que celles qui avaient seulement un ou aucun leurre mal construit. Ainsi, les questions contenant 3 leurres mal construits avaient un indice de difficulté variant de 0,71 à 0,94 pour les sept examens. La majorité des étudiants obtenaient ainsi la bonne réponse à ces questions qui avaient au moins un choix de réponses mal formulé. La faible proportion (12,3 %) de questions avec aucun leurre mal construit s'explique, selon les auteurs, par le fait que les questions ont été rédigées par des professeurs internes au programme de soins infirmiers, dont la plupart ont une formation minimale en rédaction de questions à choix multiples. Ce manque d'expérience ou de connaissances en rédaction de questions peut expliquer pourquoi les questions avec des leurres mal construits étaient plus faciles. Les étudiants ont peut-être pu déduire la bonne réponse en fonction des options présentées, les leurres mal construits n'étaient peut-être pas plausibles selon le contexte de la question.

Il va sans dire que l'inclusion de leurres mal construits dans les questions d'examens à des impacts sur la qualité des évaluations. Les questions avec des leurres mal construits peuvent ajouter une difficulté artificielle aux résultats de l'évaluation, elles deviennent alors dans certains cas plus faciles (Downing, 2002). Cette situation interfère avec l'interprétation exacte et significative des résultats des évaluations et cela a un impact sur le taux de réussite des étudiants, en particulier pour ceux qui sont près du seuil de passage. Pour s'assurer de la qualité des évaluations, des analyses d'items sont effectuées. Or, cela peut mener à l'élimination de questions a posteriori (Downing, 2002; Haladyna *et al*., 2002). Cette façon de faire affecte l'interprétation possible des résultats, ou dit autrement, réduit les preuves reliées au construit de l'examen, car moins de questions entrent en ligne de compte. La validité de construit selon Downing (2002) « se réfère à de multiples sources de

données probantes appuyant ou réfutant les interprétations significatives et précises des résultats des tests, les déductions tirées des évaluations et les décisions du seuil de passage prises à l'égard des examinés à partir de ces évaluations » (p.S103). De plus, lorsque des leurres mal construits font partie des questions retenues dans l'évaluation, l'interprétation des scores peut être erronée comme nous le démontrent les études ci-haut rapportées.

2.4.3.2. Nombre optimal d'options

Une seconde consigne de rédaction pour laquelle nous avons des données probantes porte sur le nombre optimal d'options à inclure dans les questions à choix multiple. Dans l'étude de Schneid et collaborateurs (2014), ils ont observé, en termes de temps et de propriétés psychométriques, l'effet de passer de quatre ou cinq options à trois options de réponses. Pour ce faire, ils ont administré deux versions d'un examen contenant 98 QCMs à 38 étudiants de deuxième année de médecine et à 39 étudiants de troisième année en pharmacie. Les questions avec quatre ou cinq options de réponses ont été converties en question à trois choix de réponses (deuxième version de l'examen). Les auteurs ont conclu que les étudiants répondent plus rapidement aux questions avec trois options de réponses que celles avec cinq options. Ils épargneraient en moyenne cinq secondes par QCM. Cette différence de temps permettrait aux étudiants de répondre à plus de questions pour un même temps donné. Cela représente une augmentation de 16 % du nombre de questions à trois choix de réponses que les étudiants pourraient compléter par heure. Cela a comme effet d'augmenter la validité de contenu et d'augmenter également la fiabilité des scores. En effet, en augmentant le nombre de questions, un plus grand contenu et une plus grande variété de notions peuvent être évalués dans une même période de temps déterminée. Schnied et collaborateurs (2014) ne rapporte pas de différence significative de la difficulté et de la discrimination entre la version d'examens avec des questions à trois de réponses et la version à quatre ou cinq choix de réponses. Ainsi, les questions à trois choix de réponses pourraient être une solution à adopter si la rédaction de plusieurs leurres plausibles est très difficile pour le rédacteur.

Delgado et Prieto (1998) se sont intéressés à voir si la discrimination se voyait modifier suite à la délétion d'une option dans les questions à choix multiples. En se servant de trois

examens distincts répondus par des étudiants en médecine, une délétion du quatrième choix dans les questions de ces examens n'a pas modifié de manière significative l'indice de discrimination. Cela a eu comme effet de diminuer légèrement la difficulté sans impact sur la discrimination ou la fidélité des trois examens. Les indices de discrimination des examens avec trois ou quatre options de réponse se situaient près de 0,30, ce qui, selon la classification de Ebel et Frisbee (1991) est considéré comme bon (>0,30). Un avantage marqué dans l'élaboration de questions à 3 options de réponses en est un principalement d'économie de temps et d'énergie pour le rédacteur de questions. En effet, écrire des options plus plausibles devient plus facile ou encore les évaluations peuvent être faites dans un plus court laps de temps. Ce qui vient appuyer les résultats de l'étude de Schneid et collaborateurs (2014).

Malgré ces résultats favorables à la rédaction de questions comportant seulement trois choix de réponses, un facteur doit tout de même être considéré, soit le contexte dans lequel se déroule l'évaluation. Ainsi, si l'évaluation en est une avec un enjeu élevé (ex. : certification) il vaut peut-être mieux rédiger des questions avec quatre choix de réponses pour éviter les bonnes réponses attribuables à la chance. Avec seulement trois choix de réponse, un étudiant a ainsi 33,33 % de chance d'obtenir une bonne réponse en répondant aléatoirement à cette question. Ce pourcentage diminue à 25 % pour des questions à quatre choix de réponse

2.4.3.3. L'utilisation de l'option « Aucune de ces réponses »

L'utilisation ou l'abolition de l'option « Aucune de ces réponses » dans les choix de réponses est la troisième consigne pour laquelle des données empiriques existent dans la littérature scientifique. Un grand nombre d'examens à choix multiple comportent l'option « Aucune de ces réponses » quel que soit le domaine d'études. En pédagogie des sciences de la santé, il est toutefois quasi unanime que l'utilisation de l'option « Aucune de ces réponses » devrait être réduite, voire éliminée (DiBattista *et al.*, 2014). L'étude de Haladyna et Downing (2002) suggère même que les rédacteurs novices ne devraient pas utiliser ce choix d'options dans leur question.

L'utilisation de l'option « Aucune de ces réponses » fait référence à des situations complexes pour les étudiants. Les étudiants réagissent différemment face à ce choix, ils peuvent penser qu'il s'agit simplement d'un leurre et donc ne le considère pas davantage alors que cela pourrait tout de même être la bonne réponse. Ces situations peuvent avantager ou désavantager certains étudiants non pas en termes de connaissances, mais plutôt de la manière dont ils réagissent. Ce choix de réponse pourrait être d'autant plus difficile lorsqu'il est utilisé comme bonne réponse plutôt que comme leurre (DiBattista *et al.,* 2014). Pour arriver à prendre une décision, l'étudiant (rusé) peut utiliser une procédure de jugement relatif qui consiste à déterminer laquelle des options ressemble le plus à la bonne réponse de la question. Par conséquent, lorsque « aucune de ces réponses » est la bonne option, on peut s'attendre à ce que les candidats rusés montrent une forte tendance à sélectionner de manière erronée un leurre comme bonne réponse; inversement, on s'attend à ce qu'ils montent une faible tendance à sélectionner correctement l'option « Aucune de ces réponses ».

Les résultats de l'étude de DiBattista et collaborateurs (2014) arrivent à ces conclusions. De plus, l'étude a démontré que cette option de réponse fait en sorte que la question semble avoir un plus grand pouvoir discriminant qu'elle ne l'est réellement. En effet, leur analyse initiale de la discrimination a révélé que l'utilisation du « Aucune de ces réponses » comme option n'a pas eu d'effet statistiquement significatif sur le pouvoir discriminatoire des questions. Cependant grâce à la méthode de recherche utilisée, ils ont pu regarder un peu plus loin le phénomène. Cinq versions de 40 questions de connaissances générales ont été créées. Chacune des questions avait quatre options de réponses, dont une bonne réponse. Pour chaque version, soit la bonne réponse soit une des trois autres options a été remplacée par l'option « Aucune de ces réponses ». À côté de chaque option « Aucune de ces réponses », une ligne était mise pour que l'étudiant puisse écrire la bonne réponse, le cas échéant. Ils ont remarqué par contre que lorsque les étudiants choisissaient correctement l'option « Aucune de ces réponses », ils n'écrivaient pas la bonne réponse dans 19 % des cas, démontrant ainsi qu'ils ne connaissaient pas réellement la bonne réponse à la question posée.

Une dernière étude (Caldwell et Pate, 2013) sur le sujet nous indique également que l'utilisation du « Aucune de ces réponses » augmente la difficulté de la question seulement s'il s'agit de la bonne réponse. L'étudiant doit être en mesure de reconnaitre que les autres options ne sont pas les bonnes réponses, mais sans nécessairement savoir quelle est la véritable bonne réponse. Ainsi, l'évaluation sert habituellement à savoir quelles sont les connaissances acquises des étudiants, l'utilisation de cette option de réponse n'est donc pas appropriée. L'utilisation de l'option « Aucune de ces réponses » ne devrait pas être utilisée dans la rédaction de questions à choix multiples.

2.4.3.4. Format des questions

Finalement, la dernière consigne de rédaction de questions d'examens écrits pour laquelle il existe des données quant à l'impact du respect de cette consigne sur la qualité des questions est un peu plus générale et concerne cette fois le format des questions. Ce qui est entendu par le format des questions peut vouloir dire plusieurs choses notamment en faisant référence à l'utilisation d'un vocabulaire simple et adapté au niveau des étudiants. Le langage simplifié est un moyen efficace de réduire l'influence de la capacité de lecture, une source de variance non pertinente lorsque l'évaluation mesure autre chose. Des options de réponses qui sont homogènes font également partie d'un format standard de question, de cette façon cela aide à mettre l'accent sur l'idée principale de la question et ainsi augmenter la discrimination de celle-ci. La longueur de l'énoncé (ni trop longue ni trop courte) ainsi que la clarté de la question doivent aussi être prises en considération lors de la rédaction de question. Ainsi, une question de format standard répond à plusieurs consignes de rédaction à la fois. Cependant, lorsque des questions dites non standards se retrouvent dans les évaluations, cela affecte la performance des étudiants en modifiant la difficulté ou la discrimination de la question (Haladyna et Downing, 2002).

En résumé, la recension des écrits nous permet de constater que la rédaction de question d'examen est une tâche complexe qui nécessite une quantité importante de ressources de la part des rédacteurs (Caldwell et Pate, 2013 ; Haladyna *et al.,* 2002). La difficulté de cette tâche réside, entre autres, dans le fait qu'il existe un grand nombre de consignes de rédaction dans la littérature scientifique. Malgré l'existence de ces consignes, les

rédacteurs de questions d'examens écrits s'en servent très peu pour les aider à élaborer les évaluations. Ainsi, nous nous sommes demandé quelles sont les stratégies portant sur les consignes de rédaction qui existent pour assurer la qualité des évaluations dans le domaine des sciences de la santé, notamment en médecine.

Troisième chapitre : OBJECTIFS SPÉCIFIQUES DE RECHERCHE

En considérant la problématique abordée au sujet de la rédaction de questions d'examens écrits et de l'utilisation de consignes de rédaction ainsi que de l'état des connaissances issues de la recension des écrits, cette étude poursuit trois objectifs de recherche.

Ces objectifs spécifiques sont donc :

1) Identifier les consignes de rédaction discriminantes

2) Qualifier l'utilisation et l'acceptabilité des consignes auprès des rédacteurs

3) Quantifier l'impact de leur utilisation sur les propriétés psychométriques des questions lors de la rédaction d'examens écrits

La présentation du premier article permettra de répondre à l'objectif un (1) tandis que le deuxième article qui sera présenté plus loin répondra aux objectifs deux (2) et trois (3).

Quatrième chapitre : Articles scientifiques

Article 1

L'identification de consignes discriminantes pour la rédaction d'examens écrits

Auteurs de l'article : Élise Vachon Lachiver, Christina St-Onge, Jacinthe Cloutier et Paul Farand

Statut de l'article : L'article a été soumis dans la revue scientifique Pédagogie Médicale en août 2016, il demeure sous-évaluation depuis cette soumission.

Avant-propos

Les analyses statistiques, l'interprétation des résultats ainsi que la rédaction de l'article scientifique ont été réalisées par Élise Vachon Lachiver (première auteure). L'article a été révisé par la directrice de recherche (CSTO) ainsi que les coauteurs (JC et PF). La directrice de recherche a également encadré l'ensemble du processus de recherche et de rédaction. Une autorisation a été reçue par tous les coauteurs de l'article pour permettre la publication de ce dernier dans le présent mémoire.

Résumé :

Contexte/objectif : L'évaluation des apprentissages a une importance indéniable en pédagogie des sciences de la santé. L'élaboration de questions écrites est toutefois un grand défi bien qu'on retrouve dans la littérature scientifique plus de 150 consignes guidant la rédaction. Malheureusement, il existe peu de données probantes quant à l'impact de l'utilisation des consignes sur la qualité des questions. L'objectif de cette étude est d'identifier les consignes de rédaction de questions écrites qui permettent de différencier les bonnes des moins bonnes questions, et donc qui devraient être respectées lors de la rédaction de celles-ci.

Méthode : Nous avons sélectionné 36 consignes de rédaction parmi plus de 150 répertoriées dans les écrits scientifiques. Nous les avons utilisés pour évaluer 407 questions : 150 questions à choix multiples (QCM) et 141 questions à réponse ouverte courte (QROC) répondues par les étudiants de 1re et 2e années en médecine ainsi que 116 questions QCM répondues par 325 étudiants de 4e année en médecine. Une analyse

d'items, basée sur les principes de la théorie classique des tests, a été réalisée afin d'identifier les consignes de rédaction les plus discriminantes.

Résultats : Six consignes de rédaction discriminantes ont été identifiées pour les QROC (discrimination moyenne = 0,20) et huit consignes de rédaction discriminantes ont été identifiées pour les QCM (discrimination moyenne = 0,19; accord interjuges moyen = 0,98).

Conclusions : Nos résultats suggèrent que certaines consignes de rédaction de questions d'examen peuvent discriminer entre les bonnes et moins bonnes questions, ce qui pourrait aider à l'élaboration d'examens de plus haute qualité. Cette hypothèse devra faire l'objet d'études futures.

Mots clés : Évaluation, Examens écrits, Rédacteurs de questions, Propriétés psychométriques, Étudiants en médecine

Abstract :

Aim : The assessment of learning is undeniably important in health sciences education. The development of written questions is a real challenge for the writers. The scientific literature has more than 150 instructions guiding the writing for written exam questions. Unfortunately, there is little evidence about the impact of respecting or not those guidelines on the quality of questions. The objective of this study is to identify the guidelines that differentiate good and bad exam questions, so that they should be respected when writing questions.

Method : We selected 36 guidelines of the 150 listed in the literature. We used those guidelines to assess 407 written questions exams: 150 multiple choice questions (MCQ) and 141 short-answer questions (SAQ) answered by students of 1st- and 2nd-year of medicine and 116 questions answered by 325 4th-year medical students. An items analysis, based on the principles of the Classical Test Theory was performed to identify the guidelines that are the most discriminant.

Results : We identified six discriminant guidelines for SAQ (mean discrimination = 0.20) and eight discriminant guidelines for MCQ (mean discrimination = 0.19; mean inter-rater agreement = 0.98).

Conclusion : Our results suggest that some of the guidelines for written exams can discriminate between good and bad questions, which could help to develop more high-quality evaluations. This hypothesis should be tested in future studies.

Key words : Evaluation, Written exams, Exam developers, Psychometric properties, Medical student

Introduction

L'évaluation des apprentissages et des compétences est au cœur de la pédagogie des sciences de la santé, tant au plan préclinique qu'à l'externat ou à la résidence[1]. Les évaluations sont utilisées notamment au préclinique, pour assurer un premier point de contrôle auprès des étudiants, c'est-à-dire que les évaluations permettent de vérifier que les étudiants possèdent les connaissances adéquates et les habiletés nécessaires pour intégrer la clinique, et ainsi, prendre en charge des soins accordés aux patients[2-5]. Dans ce cas, l'évaluation peut avoir d'importantes conséquences, telles que l'accès à une profession souhaitée[3]. Les programmes de médecine veulent aussi pouvoir identifier les étudiants en difficulté par le biais des évaluations bien construites. Ces dernières permettraient de discriminer les étudiants plus faibles qui pourraient se retrouver en situation d'échec[6]. Pour répondre à ces différents objectifs de la formation médicale, le développement d'examens doit être fait de manière rigoureuse et appropriée[1] et rencontrer les plus hauts standards de qualité.

L'élaboration de questions d'examens écrits est souvent décrite comme étant autant un art qu'une science[7-8]. Elle est souvent considérée comme une tâche ardue[9-11] et elle représente un défi de taille. En effet, la rédaction de questions d'examens écrits demande beaucoup de temps, d'efforts, de ressources et de connaissances de la part des rédacteurs de questions[12-13]. Malheureusement, il arrive que des questions d'examens soient mal rédigées (problèmes de contenu, ambigüités, faibles propriétés psychométriques, etc.) et doivent être retirées de l'évaluation lors de l'analyse psychométrique. En modifiant le contenu de l'évaluation *a posteriori*, des conséquences négatives sont alors présentes. Cela vient diminuer la validité de l'évaluation, et ce, sur plusieurs plans[5,14-16]. La validité selon Bertrand et Blais[17] « consiste en un jugement basé sur des preuves empiriques et sur une argumentation de nature théorique qui vise à justifier l'interprétation des scores obtenus à la suite de l'administration d'un test dans un contexte donné » (p.240). Par exemple, lors de l'élaboration d'un examen, les rédacteurs doivent habituellement élaborer des questions représentatives d'un certain contenu. Ainsi, lorsque des items sont retirés de l'examen, le contenu est modifié et n'est peut-être plus entièrement représentatif du contenu devant être évalué. Cette modification au plan de la représentativité du contenu peut s'avérer problématique, car

de l'information considérée initialement comme étant importante manque dans le jugement évaluatif.

Pour les rédacteurs, une option quant à la rédaction de questions de qualité est de se tourner vers les écrits scientifiques où l'on retrouve une panoplie de consignes de rédaction[1,13-16,18]. Toutefois, avec plus de 150 consignes répertoriées, il peut être difficile pour les rédacteurs de s'y retrouver faute d'expertise et de ressources. De plus, il existe peu de données empiriques quant à ces consignes, c'est-à-dire, qu'il y a peu de données nous indiquant si le respect, ou non-respect, de ces consignes a un impact sur les propriétés psychométriques des questions d'examens écrits. Il est ainsi encore plus difficile pour les rédacteurs de naviguer à travers cette panoplie de consignes recensées dans les écrits scientifiques. Ainsi, l'objectif de cette étude est d'identifier les consignes de rédaction de questions qui permettent de discriminer les questions d'examens écrits de haute et de basse qualité.

Méthodologie

Devis

Une analyse d'items, basée sur la théorie classique des tests, a été réalisée dans le cadre de cette étude afin de nous permettre d'obtenir les indices de discrimination et de difficulté pour chaque consigne étudiée. Les résultats de l'analyse d'items ont été utilisés pour identifier les consignes de rédaction de questions d'examens écrits qui permettent de différencier les questions de qualité de celles de moins bonne qualité.

Matériel

Une liste de consignes de rédaction a été élaborée à partir d'une recension des écrits scientifiques. Les mots clés utilisés étaient « item writing », « rules for writing MCQ » et « quality MCQ », et les recherches ont été effectuées à partir des moteurs PubMed, ERIC et Google Scholar. Au total, 154 règles de rédaction ont été répertoriées dans sept articles scientifiques et un livre (références en annexe). Il est à noter que ces articles comportaient, pour la plupart, des recensions exhaustives de manuels spécialisés sur l'évaluation des apprentissages ou des compétences. Le seul livre retenu n'avait pas été inclus dans les articles recensés (puisque rédigé en français).

Un processus itératif de sélection des consignes, illustré à la Figure 1, a été mis en place pour sélectionner les consignes de rédaction à retenir. Dans un premier temps, seules les consignes portant sur la rédaction (et non sur l'administration ou les aspects logistiques) des questions à choix multiples (QCM) et des questions à réponses ouvertes courtes (QROC) ont été retenues (retrait de 14 consignes de rédaction). La sélection de ces types de questions (QROC et QCM), s'est fait puisqu'ils sont largement répandus dans les examens écrits. Dans un deuxième temps, ces consignes ont été révisées de façon itérative afin d'obtenir une grille d'énoncés opérationnels (consigne claire et applicable; retrait d'une consigne de rédaction), mutuellement exclusifs (évaluant des éléments distincts; retrait de 89 consignes de rédaction) et adaptés au contexte d'évaluation en pédagogie des sciences de la santé (retrait de 15 consignes de rédaction).

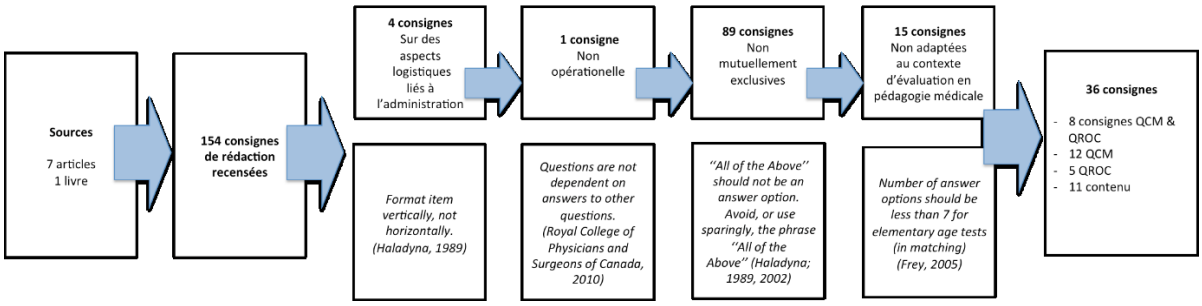


Figure 1. Identification d'un bassin de consignes de rédaction appropriées pour les examens en pédagogie des sciences de la santé.

Au terme de ce processus, 35 consignes ont été retenues, transformées en énoncés et regroupées en quatre catégories : (a) non différenciée (c.-à-d. pouvant s'appliquer aux QCM et QROC); (b) QCM; (c) QROC et (d) contenu (c.-à-d. qui requièrent une expertise de la matière pour être évalués). La liste de consignes a été prétestée par un expert de forme et par un expert de contenu sur une trentaine de questions. Après cette première mise à l'essai, quelques correctifs ont été apportés pour assurer une compréhension uniforme et, par conséquent, une utilisation standardisée des consignes. L'échelle finale comportait 36 consignes (un énoncé de la grille préliminaire a été scindé en deux), réparties en quatre catégories. La catégorie « non différenciée » contenait 8 consignes, la catégorie « QCM » en contenait 12, la catégorie « QROC » en contenait cinq et la catégorie « contenu » contenait 11 consignes.

Données

Trois banques de questions ont fait office de bases de données (voir leur description au Tableau 1). Un total de 407 questions différentes a été considéré. Les propriétés psychométriques des questions d'examens utilisées dans cette étude sont résumées au Tableau 1. Aucune donnée concernant les étudiants n'a été utilisée lors de cette étude, seules les données de groupe ont servi à mesurer les indices de discrimination; ainsi, aucune autorisation d'un comité d'éthique n'a été nécessaire selon la politique d'éthique des trois conseils[19].

Tableau 1 : Description des données

Caractéristiques	Banque de questions 1	Banque de questions 2	Banque de questions 3
Nombre de questions	141	150	116
Type de question	QROC	QCM	QCM
Provenance	Banque de questions du programme préclinique de médecine UdeS	Banque de questions du programme préclinique de médecine UdeS	Données d'un projet de recherche précédant (influence du testing)[20]
Niveau testé	200 étudiants 1ère et 2ième année en médecine de l'UdeS	200 étudiants 1ère et 2ième année en médecine de l'UdeS	325 étudiants en médecine des 4 universités du Québec
Indices mesurés	-Indice de discrimination	-Indice de discrimination -Accord inter-juge	-Indice de discrimination -Accord inter-juge

Banque de questions 1

Cent quarante et un (141) QROC ont été choisies aléatoirement de la banque de questions du programme préclinique de médecine de l'Université de Sherbrooke (UdeS). Plus spécifiquement, les questions provenaient des examens écrits utilisés dans 15 unités du programme d'apprentissage par problèmes (Introduction au programme MD, Biomédical I et II, Médecine et santé des populations, Système nerveux, Psychiatrie, Système musculo-squelettique, Santé publique, Cardiologie, Système respiratoire, Système digestif, Urologie, Hématologie, Maladies infectieuses, Endocrinologie et Système oto-rhino laryngologique). Elles ont été répondues par une cohorte d'environ

200 étudiants de médecine, qui étaient en première ou deuxième année. Les questions ainsi que leurs propriétés psychométriques (indice de discrimination et de difficulté) ont été obtenues par le biais du programme de médecine. L'indice de difficulté représente la performance moyenne d'une question alors que l'indice de discrimination représente la capacité de la question à faire la distinction entre les étudiants forts et les étudiants plus faibles.

Banque de questions 2

Cent cinquante (150) QCM ont également été choisies aléatoirement dans la banque de questions du programme préclinique de médecine de l'Université de Sherbrooke. Ces questions avaient été répondues par environ 200 étudiants et provenaient des mêmes unités d'enseignement susmentionnées. Les indices de difficulté et de discrimination ont été recueillis auprès du programme pour chacune des questions retenues.

Banque de questions 3

Une troisième banque de questions a également été utilisée lors de l'étude. Il s'agit de 116 QCM avec vignettes cliniques répondues par 325 participants dans le cadre d'une étude sur l'influence du testing[20]. Les propriétés psychométriques (indices de discrimination et de difficulté) de ces questions ont été obtenues auprès de l'équipe de recherche.

Procédure

Pour chacune des 407 questions décrites ci-haut, nous avons déterminé si elles respectaient (ou non) chacune des 36 consignes de rédaction retenues dans chacune des banques de données respectivement. Ces données (0 et 1) ont été utilisées dans le cadre de l'analyse d'items[17]. Un juge pour la Banque 1, un pour la Banque 2 et deux juges pour la Banque 3 ont évalué la qualité des questions de chacune des banques respectivement. Les données provenant d'une autre étude, les juges ont été sélectionnés préalablement sans que nous ayons mot à dire.

Analyses

Dans le but de vérifier si les consignes retenues permettent d'avoir une idée de la qualité des questions, nous avons réalisé une analyse d'items, selon les principes de la théorie classique des tests (Crocker et Algina, 1986) pour vérifier la discrimination de chacune

des consignes répertoriées. Les analyses sont résumées à la Figure 2. L'indice de discrimination (corrélation entre le score à l'item et le score total corrigé (score total – le score à l'item)) a été calculé pour chaque consigne, et ce, à partir des questions incluses dans chaque base de données. Pour la Banque 3, l'accord inter-juges (pourcentage d'accord) a été fait sur chaque consigne pour vérifier la standardisation quant à l'évaluation du respect des consignes de rédaction de questions entre deux juges pour cette banque de données. Le respect des consignes portant sur le contenu a été évalué par deux résidents médecins. Le respect des autres consignes a été évalué par deux étudiants gradués en mesure et évaluation. Il s'agit de la seule banque de questions pour laquelle deux juges évaluaient la même question, il a alors été possible de considérer l'accord inter-juge pour les questions de cette banque uniquement. Les données ont été colligées dans un fichier Excel et exportées dans SPSS 17.0 pour les analyses d'items. L'accord inter-juges a été calculé dans Excel version 14.

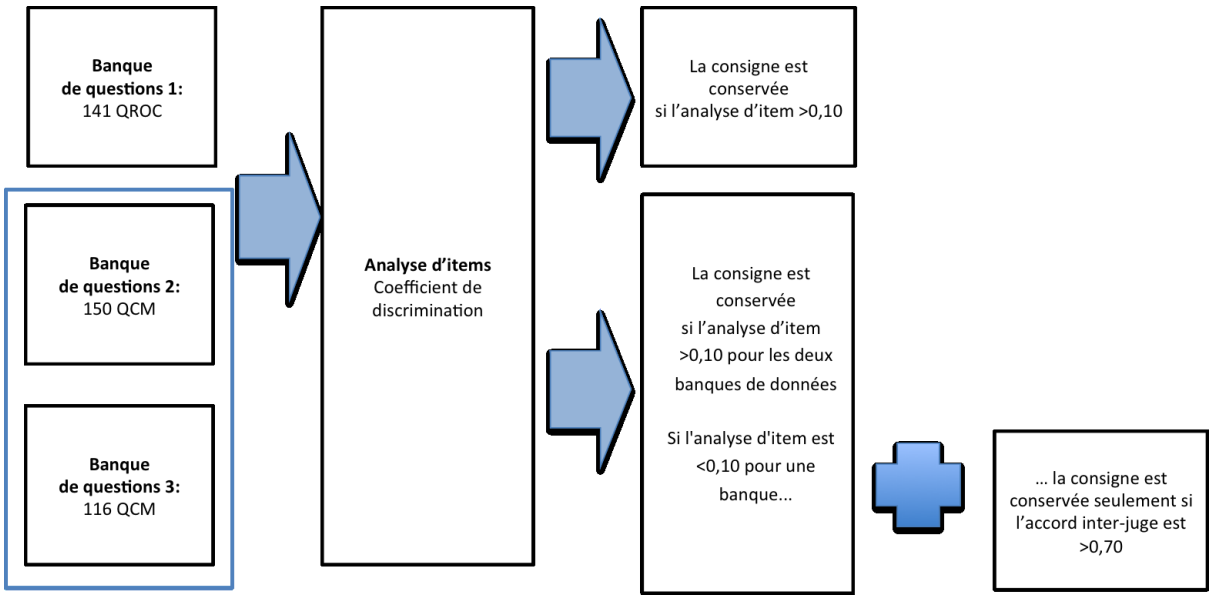


Figure 2. Sélection des consignes les plus discriminantes

Sur la base de ces résultats, nous avons déterminé, pour chaque consigne, si elle permettait (ou non) de discriminer les questions d'examens écrits de haute et de faible qualité. Les analyses ont été faites sur les trois banques de questions séparément puisqu'une des banques ne comportait que des QROC alors que les deux autres banques de QCM provenaient de projets différents (certaines données différaient donc). Pour la Banque 1 : le critère pour conserver une consigne était d'obtenir un coefficient de discrimination supérieur à 0,10 une fois l'analyse d'item effectuée. Pour les Banques 2

et 3, les critères d'inclusion étaient : 1) deux coefficients de discrimination supérieurs à 0,10 (pour une consigne) ou 2) un coefficient de discrimination supérieur à 0,10 et l'accord inter-juges supérieur à 0,70 (classification selon Ebel et Frisbie (1993)). Cette dernière stratégie a été choisie, car cela représente une source d'informations supplémentaire puisqu'il existe certaines contradictions dans les deux banques de données relatives aux QCM. Nous estimons donc que, si la consigne est comprise par les juges, elle le sera probablement par les rédacteurs de questions. De plus, une discrimination négative a été considérée comme un critère d'exclusion. En effet, les consignes ayant un coefficient de discrimination négatif ne permettaient pas de différencier les questions de bonne qualité de celles ayant une moins bonne qualité. C'est pourquoi ces consignes n'ont pas été retenues.

Résultats

Le Tableau 2 contient les moyennes des indices de discrimination, les moyennes d'accord inter-juges de chacune des consignes de rédaction pour chaque banque de questions en plus d'indiquer s'il s'agit d'une consigne faisant référence au type de questions (tous les types de questions, QCM, QROC) ou au contenu.

Tableau 2 : Indices de discrimination et accord inter-juges des consignes de rédaction selon les banques de questions.

Consignes de rédaction	Banque de questions 1 <QROC>	Banque de questions 2 <QCM>	Banque de questions 3 <QCM>	
	Indice de discrimination	Indice de discrimination	Indice de discrimination	Accord inter-juges
Applicables à tous les types de questions				
1. La question d'introduction doit être une phrase complète.	0,042	0,074	0,042	1,00
2. L'énoncé de la question doit porter sur un seul problème, un seul contenu.	0,042	-0,032	-	1,00
3. La question d'introduction doit être brève.	-0,015	0,117‡	-0,045	0,95
4. Utiliser un vocabulaire approprié pour le niveau du répondant.	**0,263**†	-0,062	-	1,00
5. Les directives dans l'énoncé de la question sont claires et le répondant comprend exactement ce qui est demandé.	0,018	0,181‡	0,243‡	**1,00**†
6. Les directives doivent indiquer clairement s'il faut identifier la bonne ou la meilleure réponse.	-	-0,273	0,318‡	0,41
7. Éviter les mots inutiles dans l'énoncé de la question.	-	-0,003	0,134‡	0,98
8. Ne pas utiliser de formulation négative.	**0,161**†	0,026	-0,190	0,95
Applicables aux QCM				
9. Utiliser les questions à choix multiple pour mesurer des niveaux de pensée plus élevés. (Il faut éviter de faire des questions de type 'rappel' de connaissances)	N/A	0,191‡	-0,181	0,64
10. Éviter les choix de réponse pouvant susciter du "test-wiseness". Par exemple, éviter les choix de réponse absurdes, invitants (formal prompts) ou les indices sémantiques (trop spécifiques/généraux).	N/A	0,070	0,164‡	**0,95**†
11. Inclure l'idée principale et la majeure partie de la phrase dans l'énoncé de la question.	N/A	0,318‡	-0,024	1,00
12. Éviter les questions de Type K (choix de réponse complexes, ex.: a et b, mais pas c).	N/A	-0,003	-	1,00
13. ‘‘Je ne sais pas’’, ‘‘Aucune de ces réponses’’ ou ‘‘Toutes ces réponses’’ ne doivent pas être un choix de réponse.	N/A	0,195‡	0,229‡	**1,00**†
14. La longueur des choix de réponse doit être constante.	N/A	0,067	-0,070	0,79
15. Les choix de réponse ne doivent pas être plus longs que l'énoncé de la question.	N/A	-0,254	-0,089	0,93
16. Les choix de réponse ne doivent pas contenir de mots répétitifs lorsque possible.	N/A	-0,093	0,293‡	1,00
17. Les choix de réponse doivent être mutuellement indépendants/exclusifs.	N/A	0,330‡	0,208‡	**0,95**†
18. Le contenu des choix de réponse doit demeurer homogène.	N/A	-	-	0,95

Consignes de rédaction	Banque de questions 1 <QROC>	Banque de questions 2 <QCM>	Banque de questions 3 <QCM>	
	Indice de discrimination	Indice de discrimination	Indice de discrimination	Accord inter-juges
19. Éviter l'utilisation d'adverbes tels que parfois, quelques fois, toujours, jamais, etc.	N/A	-	-	1,00
20. Éviter de pister la bonne réponse par des constructions grammaticales erronées.	N/A	0,155‡	0,229‡	**0,98**†
Applicables aux QROC				
21. Rédiger la question de telle sorte qu'il n'existe qu'une seule bonne réponse.	**0,195**†	N/A	N/A	N/A
22. Éviter les réponses qui dépassent une courte phrase.	**0,287**†	N/A	N/A	N/A
23. Éviter les réponses qui comprennent plus de 6 éléments (mot, expression, etc.).	0,022	N/A	N/A	N/A
24. Indiquer le degré de précision attendu, lorsque pertinent.	**0,145**†	N/A	N/A	N/A
25. Indiquer si l'insertion d'éléments non pertinents sera pénalisée.	**0,138**†	N/A	N/A	N/A
Applicables aux questions de contenu				
26. Utiliser des distracteurs plausibles; éviter les distracteurs illogiques.	-0,070	-	0.055	1,00
27. Créer chacune des questions à partir d'un objectif éducationnel.	-0,094	-	-	1,00
28. Les choix de réponse doivent inclure seulement une réponse correcte.	-	-	0.175‡	**1,00**†
29. Les exemples et les énoncés ne doivent pas provenir du *textbook.*	-	-	-	1,00
30. Les distracteurs devraient inclure des erreurs communément commises par les étudiants.	-	-	-	1,00
31. Éviter de rédiger des choix de réponse d'une façon trop technique.	-	-	-0.115	0,31
32. Utiliser des expressions familières qui sont incorrectes dans les distracteurs.	-	-	-	0,98
33. Utiliser des énoncés vrais qui répondent incorrectement à la question.	-	-	0.155‡	0,67*
34. Éviter des questions ambigües qui pourraient empêcher de bien répondre à la question.	-	0.117‡	-0.020	0,67
35. Éviter de développer les questions nécessitant des connaissances trop spécifiques.	-	-	0.155‡	**0,98**†
36. Éviter les questions créées à partir d'opinion.	-	-	0.175‡	**1,00**†

Légende :
† = inclus dans la grille
‡ = supérieur à 0,10 dans une banque QCM, doit vérifier l'accord inter-juges
* = exclu à cause de l'accord inter-juges plus petit que 0,70
N/A = Consigne non applicable pour ce type de question
- = Manque de données pour l'analyse complète

Consignes relatives aux QROC

Six consignes de rédaction associées aux QROC satisfaisaient le critère d'inclusion de coefficient de discrimination > 0,10. Les coefficients de discrimination obtenus sur les QROC varient de 0,138 à 0,287 et la discrimination moyenne pour ces consignes est de 0,198 (ET=0, 063) dans la Banque de questions 1. Les deux premières consignes retenues concernent des règles de rédaction s'appliquant à tous les types de questions (Consignes 4 et 8). Elles ne sont pas spécifiques pour les QROC, par contre, on remarque qu'elles sont plus discriminantes lorsqu'appliquées sur des QROC. Les quatre autres consignes de rédaction font référence à la forme d'une question à réponse ouverte spécifiquement (Consignes 21, 22, 24 et 25).

Consignes relatives aux QCM

Pour les deux banques de données (Banques 2 et 3), 16 consignes de rédaction de questions avaient, soit dans la Banque 2, soit dans la Banque 3 ou dans les deux banques en même temps, un coefficient de discrimination supérieur à 0,10. Ces consignes sont identifiées dans le Tableau 2.

Quatre consignes de rédaction ont été retenues, car leur coefficient était plus grand que 0,10 dans les deux banques de questions. Il s'agit des Consignes 5 (discrimination Banque 2=0,181; discrimination Banque 3=0,243), 13 (discrimination Banque 2=0,195; discrimination Banque 3=0,229), 17 (discrimination Banque 2=0,330; discrimination Banque 3=0,208) et 20 (discrimination Banque 2=0,155; discrimination Banque 3=0,229). De plus, l'accord inter-juges de ces consignes variait de 0,95 à 1,00. La première consigne retenue (5) s'applique à tous les types de questions alors que les trois autres (13,17 et 20) font référence particulièrement à la rédaction d'une QCM.

Quatre consignes (10, 28, 35 et 36) obtiennent un coefficient de discrimination adéquat selon les critères de sélection (>0,10), mais dans une seule des deux banques de données. Donc, ce seul critère n'est pas suffisant pour conserver la consigne. L'accord inter-juges a dû être considéré pour l'identification des consignes les plus discriminantes. Ainsi, ces quatre consignes de rédaction (nommées ci-haut) ont pu être conservées en prenant compte de l'accord inter-juges (> 0,70). Les Consignes 10 (discrimination Banque 3= 0,164), 28 (discrimination Banque 3= 0,175), 35

(discrimination Banque 3= 0,155) et 36 (discrimination Banque 3= 0,175). L'accord inter-juges pour ces consignes se situait entre 0,95 et 1,00.

Une consigne (33) a été exclue à cause du faible taux d'accord inter-juges (0,67), même si le coefficient de discrimination est supérieur à 0,10 dans la Banque de questions 3. De plus, sept consignes de rédaction (Consignes 3, 6, 7, 9, 11, 16 et 34) n'ont pas été retenues nonobstant le coefficient de discrimination (>0,10) et l'accord inter-juges (>0,70) répondant aux critères de sélection (Tableau 2). Ces consignes avaient une discrimination négative dans l'une ou l'autre des banques de questions, c'est la raison pour laquelle elles n'ont pas été retenues.

Au final, huit consignes de rédaction pour les QCM ont été identifiées comme étant discriminantes selon les critères de sélection. La discrimination moyenne des consignes sélectionnées est de 0,191 (ET=0, 031) alors que l'accord inter-juges moyen est de 0,98. L'accord inter-juges a été nécessaire pour identifier certaines consignes qui discriminent les questions d'examens écrits de haute et de faible qualité. Toutes les consignes retenues discriminaient les QCM de haute et de faible qualité. Parmi celles-ci, une s'applique à tous les types de questions, quatre portent sur la forme d'une QCM et trois font référence au contenu de la question rédigée.

Discussion

L'élaboration de questions de qualité pour les examens écrits n'est pas une activité aisée pour les rédacteurs. Une des stratégies employées pour faciliter la tâche est l'utilisation de consignes de rédaction. De par cette étude, nous avons identifié 14 consignes de rédaction qui peuvent discriminer les questions de bonne qualité des moins bonnes. Ces consignes sont toutes opérationnelles, mutuellement exclusives et applicables au contexte de l'évaluation des sciences de la santé. Elles portent sur le contenu de la question, ainsi que sur la forme que prend celle-ci, c'est-à-dire qu'elle soit à choix multiples ou à réponse ouverte courte. Cependant, certaines consignes sont plus discriminantes lorsqu'elles sont respectées dans la rédaction d'une QCM, par exemple. Ainsi, un plus grand nombre de consignes ont été retenues pour les QCM faisant référence au contenu ou encore à des consignes qui ne sont pas spécifiques à ce type de question.

Les résultats de notre étude donnent, aux rédacteurs de questions et à tous ceux qui s'intéressent à l'élaboration de bonnes questions d'examens écrits, des pistes quant à l'utilisation de certaines consignes au moment de la rédaction. Les rédacteurs de questions peuvent maintenant se baser sur ces résultats, c'est-à-dire de suivre ces consignes identifiées comme discriminant les bonnes questions au moment de la rédaction tout en structurant le processus. Sept consignes relatives aux QCM n'ont pas été retenues malgré le respect des critères de sélection que nous avions établis à la base, c'est-à-dire d'obtenir un coefficient de discrimination supérieur à 0,10. Cependant, au cours de l'analyse des résultats, il nous semblait incongru de sélectionner des consignes qui avaient une discrimination négative dans une ou l'autre des banques de questions sur les QCM.

Dans le but de rendre ces consignes facilement utilisables pour les rédacteurs, les consignes de rédaction sélectionnées ont été réparties en deux grilles de consignes, soit une pour la construction de QCM et une seconde pour les QROC. Ces grilles se veulent être un outil pratique pour les rédacteurs de façon à bien s'intégrer dans leur processus de rédaction de questions sans alourdir cette tâche complexe. De plus, les résultats de l'accord inter-juges concernant le respect (ou non) des consignes relatives aux QCM nous laissent penser qu'elles sont claires, compréhensibles et facilement applicables de la part des rédacteurs de questions. Il s'agit d'un bon indicateur d'autant plus que ces rédacteurs n'ont reçu aucune formation quant à l'utilisation de ces consignes.

Une autre retombée intéressante dans l'utilisation de ces 14 consignes de rédaction est qu'elle représente une alternative économique en terme de ressources (temps, coût et experts de contenu) contrairement à l'« *assessment engineering* » (AE) qui est une solution de plus en plus utilisée pour élaborer des examens en sciences de la santé. L'AE est un cadre de référence pour la construction de questions qui s'appuie sur les fondements de l'ingénierie[21-22]. Elle combine les connaissances et habiletés des experts de contenu avec le pouvoir algorithmique d'un programme informatique pour créer de nouvelles QCM[23]. Cependant, l'utilisation d'un programme informatique ne diminue en rien le rôle prédominant des experts de contenu dans le processus de rédaction de questions. De plus, la validité de ces questions développées par l'AE doit être vérifiée a posteriori par les experts ce qui en fait une méthode moins rapide que ce que prétendent les auteurs.

Une limite de cette étude est que les questions d'examens écrits, quoiqu'elles en étaient à leur première utilisation, étaient à la base de bonnes questions. Ces questions respectaient probablement déjà un certain nombre de consignes dû aux processus rigoureux des projets de recherche ainsi que du programme préclinique de médecine de notre institution. Ainsi, les questions qui se retrouvent dans les examens écrits sont majoritairement bien construites, créant ainsi un effet de plafond quant au respect (ou non-respect) des consignes et donc une diminution de la discrimination lors de nos analyses. Une autre limite est que nous avions seulement un juge pour les banques de questions 1 et 2. L'accord inter-juges était alors impossible à calculer. Il serait pertinent d'utiliser des banques de questions provenant d'un autre programme en sciences de la santé qui ne possède pas la même structure dans le processus d'évaluation et de création d'examens.

Conclusion

Il a été possible de réduire considérablement le nombre de consignes répertoriées dans la littérature scientifique en ciblant les consignes de rédaction qui permettent de différencier les questions de haute qualité de celles de faible qualité. Quatorze consignes de réaction (six pour les QROC et huit pour les QCM) ont ainsi été identifiées, donnant maintenant aux rédacteurs des données probantes sur lesquelles ils peuvent se baser pour rédiger leurs nouvelles questions d'examens écrits. Une étude suivra pour vérifier l'impact de l'utilisation *a priori* de ces consignes de rédaction sur les propriétés psychométriques desdites questions d'examen ainsi que sur l'acceptabilité et la faisabilité de ces grilles.

Références :

1. Case, S. M., & Swanson, D. B. (2002). Constructing Written Test Questions For the Basic and Clinical Sciences. *Director*, *27*(21), 1–181. Retrieved from http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf
2. Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, *5*(4), 238–242. http://doi.org/10.1016/j.nepr.2004.12.005
3. Epstein, R. M., & Hundert, E. M. (2002). Professional Competence. *JAMA : The Journal of the American Medical Association*, *287*(2), 226–235.
4. Linn, R. L., & Gronlund, N. E. (2000). *Measurement and Assessment in Teaching*. Merrill. Retrieved from https://books.google.com/books?id=TOZJ3ulKdi8C&pgis=1
5. Van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *BMJ (Clinical Research Ed.)*, *321*(7270), 1217–1219. http://doi.org/10.1136/bmj.321.7270.1217
6. Jouquan, J. (2002). L'évaluation des apprentissages des étudiants en formation médicale initiale. *Pédagogie Médicale*, *3*(1), 38–52. http://doi.org/10.1051/pmed:2002006
7. Downing, S. M. (2003). Validity: On the meaningful interpretation of assessment data. *Medical Education*, *37*(9), 830–837. http://doi.org/10.1046/j.1365-2923.2003.01594.x
8. Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. http://doi.org/10.1111/j.1745-3992.2005.00006.x
9. Bush, M. E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education*, *14*(4), 398–404. http://doi.org/10.1108/09684880610703974
10. Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–333. http://doi.org/10.1207/S15324818AME1503_5
11. Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, *9*, 40. http://doi.org/10.1186/1472-6920-9-40
12. Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4), 1–5. http://doi.org/10.5688/ajpe77471
13. Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143. http://doi.org/10.1007/s10459-004-4019-5
14. Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, *12*(1), 19–24. http://doi.org/10.1016/S1322-7696(08)60478-3
15. Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education*, *70*(4), 378–386.

16. Haladyna, T. M., & Downing, S. M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*. http://doi.org/10.1207/s15324818ame0201_4
17. Bertrand, R. et Blais, J-G. Modèles de Mesure: L'Apport de la Théorie des Réponses aux Items. Québec : Presse de l'Université du Québec, 2004.
18. Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New Guidelines for Developing Multiple-Choice\nItems. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*(2), 65–72. http://doi.org/10.1027/1614-2241.2.2.65
19. Conseil de recherches en sciences humaines du Canada, Conseil de recherches en sciences naturelles et en génie du Canada, Instituts de recherche en santé du Canada : *Énoncé de politique des trois Conseils : Éthique de la recherche avec des êtres humains,* décembre 2014.
20. McConnell, M. M., St-Onge, C., & Young, M. E. (2014). The benefits of testing for learning on later performance. *Advances in Health Sciences Education*, 1-16. Doi: 10.1007/s10459-014-9529-1
21. Luecht, R.M. (2009). Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm. *In D. J. Weiss (Ed.),* Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.
22. Gierl, M.J., Zhou, J., Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *Journal of Technology, Learning, and Assessment,* 7(2).
23. Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple-choice items created with automated processes. *Medical education*, 47(7), 726-733

Article 2

Utilisation *a priori* de consignes de rédaction de questions d'examens écrits en médecine : Une étude de cas multiples

Auteurs de l'article: Élise Vachon Lachiver et Christina St-Onge

Statut de l'article: L'article a été soumis dans la revue des sciences de l'éducation en mars 2017.

Avant-propos

L'étudiante a conçu le devis et le protocole de recherche ainsi que la recherche documentaire. Elle a élaboré l'outil et la collecte de données en plus d'être responsable de l'analyse et de l'interprétation des données. La rédaction de l'article scientifique a été réalisée par Élise Vachon Lachiver (première auteure). La directrice de recherche (CSTO) a révisé l'article et elle a soutenu l'étudiante durant l'ensemble du processus de recherche et de rédaction.

Résumé :

Contexte/objectif: L'évaluation des apprentissages a une grande importance dans le domaine de la pédagogie des sciences de la santé. L'utilisation d'examens écrits est omniprésente dans l'évaluation. Toutefois, l'élaboration d'évaluations de haute qualité est souvent un défi pour les rédacteurs. Il est important d'identifier des stratégies pour faciliter leur travail tout en ayant un impact positif sur la qualité des questions.

Méthode : À l'aide d'une étude de cas multiples, nous avons exploré l'utilisation d'une liste circonscrite de consignes de rédaction de questions auprès de quatre rédacteurs pour des examens d'un programme préclinique de médecine. Un questionnaire et une entrevue individuelle semi-dirigée ont été effectués pré et post utilisation des consignes. Les unités d'analyses et de comparaison étaient : propriétés psychométriques des questions pré/post utilisation des consignes, acceptabilité et utilisation.

Résultats : Pour tous les participants, les questions rédigées respectaient majoritairement les consignes de rédaction. Quoique non statistiquement significative, nous avons observé une légère augmentation du coefficient de discrimination chez les deux rédacteurs (la moitié des rédacteurs) qui utilisaient les consignes de manière active. Nos résultats suggèrent que l'appropriation que font les rédacteurs des consignes semble influencer leur perception de l'utilité et de l'acceptabilité des consignes.

Conclusions : Les résultats obtenus offrent une bonne stratégie à adopter pouvant ainsi améliorer la qualité des questions. Les rédacteurs de questions qui sont mieux outillés et qui semblent adopter pleinement l'utilisation des consignes de rédaction y voient les impacts positifs sur la qualité de leurs questions.

Mots clés : Évaluation, Examens écrits, Rédacteurs de questions, Propriétés psychométriques, Étudiants en médecine

Abstract

Aim : The assessment of trainees in health professions education has an undeniable importance and the use of written exams is omnipresent. However, developing high-quality assessment is often a challenge for exam developers. It is important to identify strategies that could facilitate the work of exam developers while having a positive impact on the quality of the written exam questions.

Method : Using a multiple case study design, we explored the use of a list of item-writing guidelines created through a rigorous and systematic process with four exam developers at the preclinical phase of an undergraduate medical program. A questionnaire was completed and a semi-directed individual interview was carried out before and after the use of guidelines. The units of analysis and comparison were: perceived acceptability and perception of the item-writing guidelines, the respect of the item-writing guideline and participants' written-exam questions' psychometric properties pre/post use of the writing guidelines.

Results : For all the participants, the written questions respected mostly the writing guidelines. Although not statistically significant, we observed a slight increase in the coefficient of discrimination among writers who actively used the instructions. Our results suggest that the participants' appropriation of the item-writing guidelines seems to influence their perception of the usefulness and acceptability of instructions.

Conclusion : The obtained results offer a new alternative for the monitoring of the quality of the questions. The exam developers who are better equipped and who seem to adhere to the use of the writing guidelines are seeing positive impacts on the quality of their questions.

Key words : Evaluation, Written exams, Exam developers, Psychometric properties, Medical student

1. Introduction et recension des écrits

L'évaluation des apprenants est une composante importante des programmes de formation (Scallon, 2004), notamment car ces derniers sont responsables de s'assurer que les étudiants acquièrent les connaissances qui leur ont été enseignées et que l'évaluation respecte les objectifs du programme (Case et Swanson, 2002) . Un enjeu important de l'évaluation des apprentissages en contexte préclinique dans les programmes de médecine est qu'elle puisse être utilisée pour identifier les étudiants n'ayant pas acquis les connaissances et les habiletés nécessaires pour passer à l'étape suivante de leur formation, soit les stages en milieu clinique (Brady, 2005; Epstein et Hundert, 2002; Linn et Gronlund, 2000; van der Vleuten, 2000).

Un monitorage de la qualité des évaluations est ainsi mis en place, mandaté par l'organisme d'agrément des programmes de médecine (CACMS Standards and Elements, 2015), pour favoriser l'utilisation de scores provenant d'examens qui satisfont à certains critères de qualité. Pour ce faire, différentes stratégies de monitorage existent, notamment l'analyse d'items (Laveault et Grégoire, 2014). Cette stratégie utilisée suite à l'administration de l'examen, permet d'identifier les questions problématiques (problèmes de contenu ou de construction, ambiguïtés, faibles propriétés psychométriques, etc.) afin de les éliminer lorsqu'il y a lieu (Downing, 2002; Haladyna *et al.*, 2002).

L'élimination des questions suite à l'utilisation de l'analyse d'items comme stratégie de monitorage peut entrainer néanmoins des conséquences négatives sur la validité de l'interprétation des résultats de l'évaluation. Bertrand et Blais (2004) décrivent la validité comme étant « *un jugement basé sur des preuves empiriques et sur une argumentation de nature théorique qui vise à justifier l'interprétation des scores obtenus à*

la suite de l'administration d'un test dans un contexte donné » (p.240). Ainsi, l'élimination de questions peut altérer le contenu de l'examen, de sorte que celui-ci ne corresponde plus entièrement aux objectifs de départ (Downing, 2003). Plus précisément, si l'évaluation contient, par exemple, cinq notions importantes et que suite à la correction toutes les questions relatives à une de ces cinq notions sont éliminées, on ne peut pas affirmer que l'interprétation du score est toujours valide, car il ne reflète plus l'entièreté du contenu devant être mesuré.

Il importe donc d'identifier des stratégies pouvant être mises en place en amont de l'administration et de la correction des questions pour en augmenter la qualité. L'utilisation de consignes lors du processus de rédaction pourrait faciliter le travail des rédacteurs tout en ayant un impact positif sur la qualité des questions écrites. Une des stratégies qui mérite d'être explorée est celle d'outiller les rédacteurs de questions au moment même où ceux-ci écrivent leurs nouvelles questions. C'est-à-dire d'utiliser *a priori* les consignes de rédactions de questions d'examens écrits et ainsi éviter de diminuer la validité de l'évaluation.

L'objectif de cette étude est donc de 1) quantifier l'impact de l'utilisation de consignes de rédaction sur les propriétés psychométriques des dites questions au moment de la rédaction d'examens écrits ; en plus de 2) qualifier l'utilisation et l'acceptabilité des consignes perçues par des rédacteurs de questions d'examens pour les unités précliniques dans un programme de médecine.

2. Méthodologie

Une étude de cas multiples a été réalisée pour répondre aux objectifs de l'étude. Selon Yin

(2009), l'étude de cas multiples est une variante de l'étude de cas simple. Elle consiste à identifier des phénomènes récurrents parmi un certain nombre de cas. En utilisant plusieurs cas lors des analyses, il est possible de s'approcher de la réplicabilité des études quantitatives. L'étude de cas multiples permet également d'identifier des convergences ou des caractéristiques entre plusieurs cas tout en permettant d'identifier certaines particularités propres à chaque cas (Yin, 2009). Il est possible de décrire les pratiques spécifiques qui sont employées dans une situation complexe comme la rédaction de nouvelles questions dans le cas présent. Finalement, la méthodologie proposée par Yin (2009) permet le recours aux données quantitatives et qualitatives, ce qui se prête bien à notre question de recherche. Pour notre étude, un « cas » réfère à un rédacteur de questions d'examens écrits. De manière plus précise, les cas étudiés représentent des rédacteurs (n = 4) de questions pour quatre différents examens du programme préclinique de médecine de l'Université de Sherbrooke (UdeS).

2.1 Participants

Tous les rédacteurs de questions d'examens écrits du programme préclinique de médecine de l'Université de Sherbrooke (N = 26) ont été invités à participer à l'étude. Les critères d'inclusion étaient : avoir rédigés des questions pour les années 2013-2014 et 2014-2015, devoir rédiger des questions pour l'année 2015-2016, mais ne pas les avoir rédigées au moment où l'étude commençait. En retournant deux ans en arrière, il sera possible de détecter les rédacteurs qui ont une amélioration naturelle de leur rédaction (amélioration de la qualité des questions) et ainsi vérifier le véritable impact de l'utilisation des consignes de rédaction des questions. Le recrutement s'est fait par courriel à tous les rédacteurs présents sur une liste qui nous a été fournie par le programme de médecine de l'Université de

Sherbrooke. Un courriel de rappel a été envoyé aux rédacteurs potentiels qui n'avaient pas répondu. Des 26 rédacteurs de questions invités, 13 ne répondaient pas aux critères de sélection : les questions étaient déjà écrites (5 rédacteurs), les rédacteurs contactés n'étaient récemment plus responsables d'un examen, c'est-à-dire qu'ils ne rédigeaient plus de nouvelles questions (8 participants). Il se peut que la liste que nous avions reçue n'était plus à jour due au fort taux de roulement des responsables de questions d'examens écrits. Ainsi, quatre rédacteurs ont accepté l'invitation et répondaient aux critères d'inclusion.

2.2 Matériel

Le guide de consignes a servi de matériel d'intervention dans cette étude tandis que la collecte de données s'est faite à partir de deux outils décrits ci-dessous.

Guides de consignes. Deux guides de consignes (Annexe 1) de rédaction de questions ont été construits sous la forme d'une grille de vérification (*checklist*) contenant des consignes de rédaction pour les questions de type réponse courte (QRO, Guide 1 : 6 consignes) ainsi que pour les questions à choix multiples (QCM, Guide 2 : 8 consignes). L'identification de ces consignes a fait l'objet d'un processus de sélection systématique et rigoureux, réalisé en plusieurs étapes, dont une recension des écrits et une analyse des propriétés psychométriques de chaque consigne retenue. Ce processus est décrit plus en détails dans Vachon Lachiver *et al.* (2016, soumis). En résumé, les 14 consignes contenues dans les guides sont celles qui permettent de différencier le mieux les bonnes des moins bonnes questions d'examens écrits.

Questionnaire d'expérience de rédaction. Un questionnaire sur l'expérience de rédaction des participants a été élaboré pour cette étude. Ce questionnaire comprend six questions qui permettent de mieux connaître les habitudes de rédaction et le niveau d'expérience des rédacteurs de questions. Plus spécifiquement, les questions portaient sur l'expérience en tant que rédacteur de questions d'examens, les stratégies utilisées lors de la rédaction, la connaissance des consignes de rédaction et la possibilité d'amélioration perçue par les rédacteurs.

Guide d'entrevue. Un guide d'entrevue a été utilisé pour documenter l'acceptabilité et l'utilité d'utiliser les consignes de rédaction, telles que perçues par les participants. Les questions contenues dans ce guide portent sur la perception des rédacteurs à propos de l'utilité des consignes, l'utilisation qu'ils en ont faite et les améliorations à apporter s'il y a lieu

2.3 Procédure

Une première rencontre de 30 minutes a eu lieu entre la chercheuse principale et les rédacteurs de questions dans le but de présenter et expliquer l'utilisation des guides de consignes de rédaction aux participants. Le formulaire de consentement (Annexe 2) ainsi que le questionnaire d'expérience de rédaction ont été remplis à ce moment par les participants. Les participants rédigeaient leurs nouvelles questions pour l'examen dont ils avaient la responsabilité au moment où ils avaient prévu de le faire, sans la présence de la chercheuse. Aucune consigne particulière ne leur a été donnée quant aux nombres et au type (QROC ou QCM) de questions à écrire. Le tout devait toutefois répondre aux

exigences du programme de médecine concernant la rédaction de questions d'examens écrits (nombre de nouvelles questions à inclure, degré de difficulté et types de question). Une seconde rencontre d'une trentaine de minutes a eu lieu deux semaines suivant la rédaction des nouvelles questions. Cette rencontre consistait en une entrevue individuelle semi-dirigée (utilisation du guide d'entrevue), conduite par la chercheuse principale auprès de chaque rédacteur. Les entrevues individuelles ont été enregistrées uniquement aux fins de transcription et d'analyse. Les indices de discriminations ont été obtenus avec la collaboration du programme de médecine suite à l'administration des examens aux étudiants.

2.4 Méthodes d'analyse des données

Données qualitatives. Les entrevues semi-dirigées ont été codées dans le logiciel *Dedoose* à l'aide d'un arbre de code élaboré de manière itérative par les coauteurs. Des analyses en silo ont d'abord été effectuées par la chercheuse principale (EVL) pour chaque cas (rédacteur) afin d'identifier les éléments reliés à certains thèmes ; l'utilisation, l'acceptabilité et la perception des consignes présentes dans les guides de rédaction. Il était important de vérifier auprès des utilisateurs des consignes comment ils les ont perçues et les raisons pour lesquelles ils les ont utilisées. De plus, il fallait voir avec eux s'ils étaient enclins à faire circuler les consignes qui leur ont été présentées dans leur milieu professionnel. Pour finir les analyses qualitatives, une analyse transversale a été effectuée afin de faire ressortir les contrastes et les caractéristiques présentes chez les quatre rédacteurs.

Données quantitatives. Pour vérifier le respect de l'utilisation de consignes de rédaction, des analyses descriptives ont d'abord été effectuées sur les données quantitatives recueillies. Chacune des questions a fait l'objet d'une analyse pour vérifier le respect ou non des consignes. Cette évaluation a été faite à l'aveuglette par l'auteure principale, c'est-à-dire qu'elle n'avait pas l'identification de l'année de rédaction à sa disposition et ne pouvait donc pas savoir si les questions avaient été écrites avant ou après avoir utilisé le guide de rédaction. Pour chaque question, un score de respect des consignes a été calculé (somme des indicateurs respectés). Des ANOVA simples ont été effectuées pour chaque rédacteur afin d'évaluer s'il y avait des différences quant aux indices de discrimination moyenne en fonction de l'année de rédaction des questions. Les analyses ont été faites sur les questions des deux années précédant l'utilisation des consignes afin de pouvoir détecter une amélioration naturelle du rédacteur à écrire de meilleures questions d'un point de vue psychométrique. Le seuil de signification a été fixé à $p = 0,05$. Les analyses ont été réalisées avec le logiciel SPSS 24.0.

2.5 Considérations éthiques

L'étude a été approuvée par le Comité d'éthique de la recherche d'éducation et sciences sociales de l'Université de Sherbrooke en juillet 2015. La confidentialité des données et l'anonymat des participants ont été respectés tout au long de l'étude.

3. Résultats

Nous présenterons dans un premier temps l'analyse des données qualitatives articulées autour de deux grands thèmes, soit l'utilité et l'acceptabilité des consignes. Par la suite, nous présenterons les résultats quantitatifs ainsi que les résultats de l'analyse transversale

(inter-rédacteurs) nous permettant d'émettre des hypothèses quant aux résultats quantitatifs obtenus.

3.1 Données Qualitatives

La moyenne d'année d'expérience de rédaction pour nos participants est de 6 ans. Le participant 2 est celui qui a le moins d'expérience (4 années comme rédacteur) alors que le participant 3 est le plus expérimenté (9 années comme rédacteur). Deux des participants nous ont dit avoir eu une formation concernant la rédaction de questions d'examens écrits (P2 et P4), les deux autres participants de notre étude ne les ont pas suivies. Des guides de rédaction de questions, ceux fournis par le programme de médecine de l'Université de Sherbrooke, sont des ressources que seulement les participants 2 et 4 utilisent lors de leur rédaction de questions.

		Participant 1	Participant 2	Participant 3	Participant 4
Utilité	**Utilisation**	Passive	Active	Passive	Active
	Motivation	Participation à l'étude	Être mieux outillé	Structurer la rédaction de questions	Formuler de meilleures questions
	Perception	Pas de surprises, consignes sont normales et correctes	Consignes concises et claires. Un guide simple et facile d'utilisation	Effet négatif à l'utilisation des consignes de manière trop systématique	Très claire réduit les risques d'erreur
Acceptabilité	**Impact perçu sur le temps de rédaction**	↓	↓	↓	↑ fait de façon plus consciencieuse
	Partage des consignes	Oui	Oui	Sous réserve	Oui

Limites à l'acceptabilité	Contexte d'évaluation ne permet pas de respecter les consignes	Consignes difficiles à appliquer dans la pratique	Évaluation en profondeur même si cela ne respecte pas toutes les consignes	Contexte d'évaluation ne permet pas de respecter les consignes

Tableau 1. Perceptions des participants quant à l'utilité et l'acceptabilité des consignes de rédaction de questions

Utilité des consignes. Deux des participants ont fait référence davantage aux consignes comme un outil pouvant structurer leur tâche. Plus précisément, le participant 1 a vu sa participation à l'étude comme étant une raison de mieux structurer sa rédaction de questions. Le participant 3, lui, a dit utiliser les consignes simplement parce qu'elles se trouvaient devant lui au moment de la rédaction. Toutefois, les deux autres participants (P2 et P4) ont indiqué, plus explicitement, utiliser l'outil afin d'améliorer la qualité de leur évaluation, voir même s'améliorer en tant que rédacteur de questions.

> *« Il [le guide] pourrait améliorer l'évaluation des apprentissages, parce que de la manière dont l'outil est fait, il permet d'être assez précis dans ce que tu vas évaluer. » Participant 2*

Acceptabilité des consignes. Nous avons regroupé sous le thème de l'acceptabilité trois sous-thèmes soit l'acceptabilité proprement dite, l'impact sur le temps de rédaction, la volonté à partager les consignes avec des collègues.

Trois des quatre participants perçoivent les consignes comme étant simples, concises et claires (Tableau 1).

> *« Un outil simple, facile d'utilisation comme ça, je n'en avais pas. Je pense que ça, c'est bienvenu. » Participant 2*

Toutefois, le participant 3 voit un effet potentiellement négatif à l'utilisation des consignes de manière trop systématique. Selon lui, le guide est davantage un conseil qu'une règle à suivre à la lettre.

> *« […] je comprends le principe, mais en même temps, ça permet d'augmenter le niveau de difficulté, je trouve, mais il faut que ce soit bien formulé pareil. C'est dans ce sens-là. En tout cas, moi, je ne verrais pas ça comme une obligation, évidemment, mais je voyais le rationnel. […] Plus on avance dans le cheminement, si l'on reproduit ce même modèle-là, pour moi, là c'est un problème. Là, on forme ce que j'appelle des robots » Participant 3*

Deux participants (P1 et P3) ont également mentionné que les consignes présentes dans le guide faisaient référence au bon sens, laissant penser qu'ils acceptent bien les consignes.

> *« […] je peux l'avoir comme guide et dire : "OK, je le regarde rapidement." Ça fait du bon sens clinique. C'est le gros bon sens, en fin de compte, quand tu regardes ça. » Participant 1*

Impact sur le temps de rédaction. Tous les participants semblaient indiquer que le temps de rédaction était demeuré identique ou était diminué, ce qui a été apprécié de tous.

> *« […] c'est sûr que ça ne raccourcit pas la rédaction, mais je ne pense pas que ça allonge de façon indue. » Participant 1*

> *« C'est sûr que ça donne une sûreté et une sécurité, une confiance. Est-ce que la confiance fait en sorte qu'on gagne du temps ? Probablement que oui. » Participant 3*

Partage des consignes. La majorité des participants, soit trois participants sur quatre, seraient enclins à partager le guide avec leurs collègues. Il est à noter que le participant ayant montré une certaine réserve quant à l'utilisation du guide par ses collègues stipule

que le guide doit être utilisé avec jugement et que les futurs utilisateurs devraient en être avertis.

> *« [je le recommanderais] sûrement, mais avec une petite note en bas de page qui dit de se servir de son bon sens. » Participant 3*

Limites à l'acceptabilité des consignes. Les participants 1, 3 et 4 ont dit ne pas respecter les consignes de rédaction dans certains contextes. Soit ils cherchaient à faire des évaluations en « profondeur », soit le contexte (sujet de l'examen) ne permettait pas de respecter toutes les consignes, notamment lorsque la question faisait référence au diagnostic différentiel. Une autre raison invoquée par les participants était que certaines difficultés rencontrées lors de la rédaction ont fait en sorte que la consigne n'a pas été respectée tout simplement.

> *« C'est sûr que c'est plus difficile, ça restreint un peu ce qu'on peut évaluer et ça fait qu'on revient un peu souvent aux mêmes choses, donc ça demande peut-être un peu d'innover. C'est peut-être le critère que, des fois, je ne suis pas tout à fait sûr si je l'ai toujours respect. » Participant 4*

3.2 Données Quantitatives

Respect des consignes. Le tableau 2 indique le nombre de nouvelles questions écrites par chaque rédacteur en fonction de l'année de rédaction. Le respect des consignes de rédaction pour les questions analysées a été quantifié dans le cadre de cette étude et ces données sont présentées dans le tableau 3. Ainsi, pour l'ensemble des questions sans distinction pour le rédacteur ou l'année de rédaction, les QROCs et les QCMs respectent, presque dans la totalité, les 14 consignes de rédaction. En effet, pour les questions à réponse ouverte courte, nous obtenons une moyenne de respect des consignes de 4,96 (ET= 0,90) sur un total de 6

consignes relatives à ce type de questions. Pour ce qui est des questions à choix multiples, la moyenne de consignes respectées était de 4,72 (ET= 0,52) sur un total de 5 consignes considérées.

Tableau 2. Nombre de nouvelles questions en fonction de l'année de rédaction

Participants	**Année de rédaction des questions**	**Nombre de nouvelles questions (n=)**
Participant 1	2013-2014	25
	2014-2015	24
	2015-2016	27
Participant 2	2013-2014	40
	2014-2015	42
	2015-2016	47
Participant 3	2013-2014	26
	2014-2015	55
	2015-2016	37
Participant 4	2013-2014	21
	2014-2015	25
	2015-2016	23

Nous avons également regardé le respect des consignes de rédaction en fonction de chaque participant plus spécifiquement. Les questions de type QCM semblent mieux respecter les consignes que celles pour les QROC pour les quatre participants. Le participant 2 est celui qui a obtenu les scores les plus bas avec 4,88 pour les questions QROCs et 4,56 pour les QCMs. Les autres participants obtiennent néanmoins des scores assez similaires soit pour le participant 1 de 4,96 pour les QROCs et de 4,74 pour les QCMs, pour le participant 3 de 5,08 pour les QROCs et de 4,74 pour les QCMs et finalement pour le participant 4 de 4,96 pour les QROCs et de 4,88 pour les QCMs.

Tableau 3 : Score du respect des consignes de rédaction

		n	Moyenne	Minimum	Maximum
Tous les participants	Question Type QROC	254	**4,96**	3,00	6,00
	Question type QCM	144	**4,72**	2,00	5,00
Participant 1	Question Type QROC	51	**4,96**	3,00	6,00
	Question type QCM	27	**4,74**	4,00	5,00
Participant 2	Question Type QROC	90	**4,88**	3,00	6,00
	Question type QCM	39	**4,56**	2,00	5,00
Participant 3	Question Type QROC	65	**5,08**	3,00	6,00
	Question type QCM	53	**4,74**	4,00	5,00
Participant 4	Question Type QROC	48	**4,96**	3,00	6,00
	Question type QCM	25	**4,88**	4,00	5,00

Coefficient de discrimination. Quoique non statistiquement significatif, nous observons une augmentation du coefficient de discrimination pour les questions de deux participants, et une diminution du coefficient de discrimination pour les questions des deux autres participants (voir Tableau 4). La discrimination moyenne pour les questions du participant 2, passe de 0,134 pour la première année à 0,187 pour la dernière année, donc on voit un gain de 0,053 à la Figure 1. Cette tendance dans l'amélioration de la discrimination moyenne des questions est plus importante entre la deuxième et troisième année (gain de 0,030) qu'entre la première et deuxième (gain de 0,023). Pour ce qui est du participant 4, la discrimination moyenne des questions augmente de 0,092 à 0,125 de la première à la troisième année, il s'agit donc d'un gain de 0,033. Encore une fois, cette tendance dans l'amélioration est plus importante entre la deuxième et la troisième année, l'augmentation de la discrimination des questions est de 0,032 (voir la Figure 1) versus 0,001 entre la première et la deuxième année. Il faut toutefois indiquer que ces différences ne sont pas significatives et ne peuvent que suggérer une tendance.

Tableau 4. Discrimination moyenne selon l'année pour chaque participant

Participant	Année de rédaction des questions	Discrimination moyenne	Écart-type
Participant 1	2013-2014	0,225	0,022
	2014-2015	0,196	0,022
	2015-2016	0,193	0,021
Participant 2	2013-2014	0,134	0,018
	2014-2015	0,157	0,018
	2015-2016	0,187	0,017
Participant 3	2013-2014	0,162	0,021
	2014-2015	0,147	0,015
	2015-2016	0,136	0,018
Participant 4	2013-2014	0,092	0,017
	2014-2015	0,093	0,015
	2015-2016	0,125	0,016

Pour les participants 1 et 3, l'indice de discrimination moyenne pour leurs questions tend à diminuer dans le temps. Ces différences sont des tendances et elles ne sont pas statistiquement significatives. Plus précisément, la discrimination moyenne des questions du participant 1 est passée de 0,225 à 0,193. La tendance de diminution de la discrimination moyenne des questions est plus importante entre la première et deuxième année soit une perte de 0,029 qu'entre la deuxième et troisième année où une légère diminution de 0,003 est survenue (voir la Figure 1). Pour le participant 3, la discrimination moyenne des questions diminue de 0,162 à 0,136. La diminution de la discrimination moyenne semble assez similaire d'une année à l'autre. La différence est de 0,015 entre la première et deuxième année et de 0,011 entre la deuxième et la troisième année. Il faut toutefois rappeler que ces différences ne sont pas significatives et ne peuvent que suggérer une tendance.

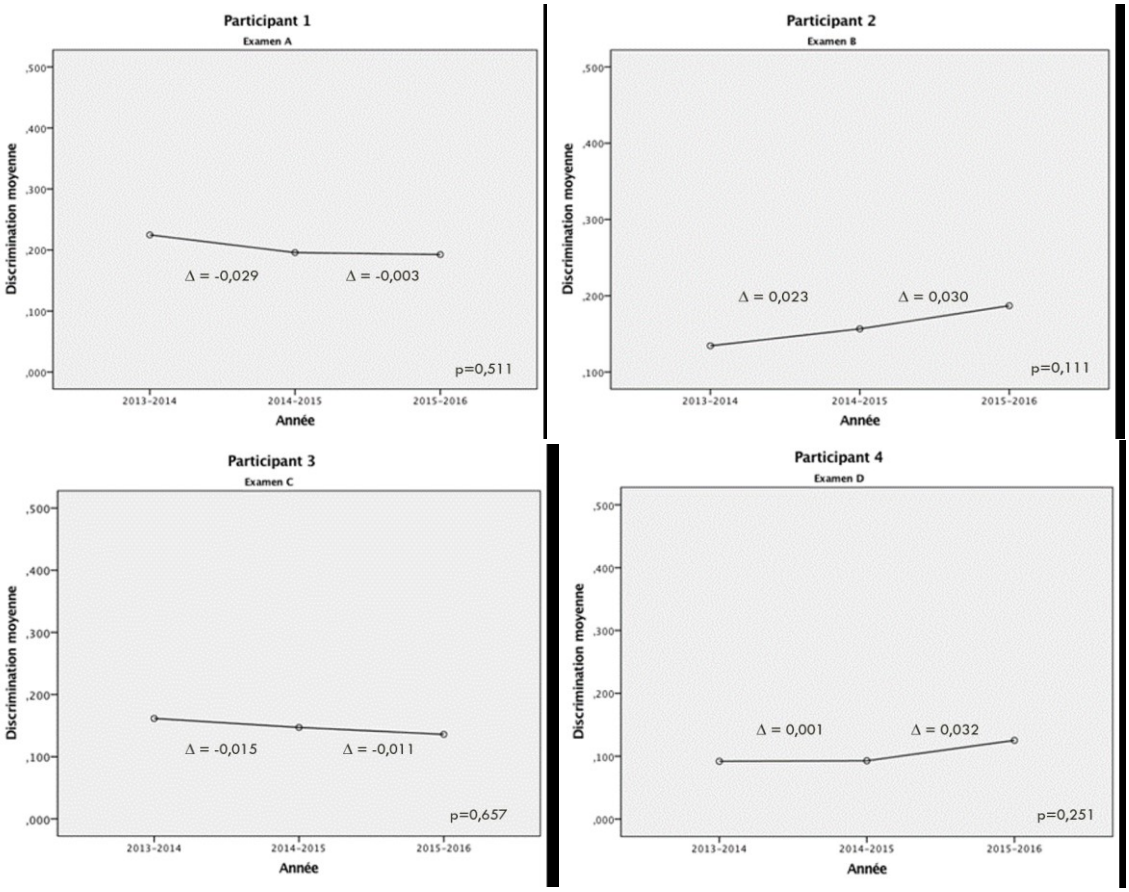


Figure 1. Discrimination moyenne des nouvelles questions pour chaque année scolaire selon le participant

En résumé, il est possible de remarquer que les participants (P2 et P4) formulent des questions qui ont des indices de discrimination plus élevés lorsqu'ils utilisent les consignes au moment de la rédaction probablement en raison des motivations qui les ont poussées à les utiliser. Tandis que les participants (P1 et P3) voient diminuer l'indice de discrimination de leurs questions malgré l'utilisation des consignes de rédaction, toutefois, cette diminution n'est pas plus prononcée que celle observée entre les deux années auparavant.

4. Discussion

La qualité des questions des examens écrits est un aspect qui est généralement contrôlé suite à la correction des examens, ce qui peut avoir comme conséquence le retrait de

questions sous-optimales. Ce processus peut entrainer une diminution de la représentativité du contenu et donc, de la validité liée à l'interprétation des scores. Nous avons donc exploré la possibilité de mettre en place une stratégie pouvant contribuer à augmenter la qualité des questions d'examens écrits *a priori*. Nos résultats suggèrent que nous avions deux types d'utilisateurs de rédaction de consignes, et que le type d'utilisation des consignes peut avoir un impact sur la qualité des propriétés psychométriques des questions.

Nous avons remarqué un bon niveau de respect des consignes par tous les participants. La plus grande différence observée entre les participants se trouve au niveau des motifs et motivations qui poussaient les rédacteurs à utiliser les consignes lors de la rédaction de questions ainsi que l'utilité perçue des consignes, ou dit autrement la perception que les participants avaient de l'outil. Cette différence au niveau de l'utilisation des consignes s'est traduite au niveau des propriétés psychométriques des questions, sans avoir de répercussion sur le respect des consignes. Plus spécifiquement, les rédacteurs qui avaient une utilisation active des consignes (deux des quatre participants) ont vu la moyenne de discrimination de leurs questions augmentée tandis que les deux autres rédacteurs de questions qui avaient une utilisation plus passive voyaient la discrimination moyenne de leurs questions diminuées. Une utilisation active des consignes de rédaction correspond à une motivation plutôt positive, de la part de ces rédacteurs, qui était à connotation plus pédagogique (apprendre à mieux évaluer), soit de formuler de meilleures questions et d'être mieux outillé.

La plus grande augmentation de coefficients de discrimination a été observée chez le participant 4. Ce participant se distingue des autres participants par sa nature réflexive quant au processus de rédaction de questions, et au choix de stratégies évaluatives. Plus spécifiquement, ce participant questionne une surutilisation des QCM et QROC, notamment pour évaluer certaines notions relatives à son domaine d'expertise. Cette constatation lui est peut-être venue parce que l'utilisation des consignes de rédaction s'est faite de manière très consciencieuse et avait comme motivation principale de formuler de meilleures questions. L'autre participant (P2) pour qui l'impact de l'utilisation des consignes a été positif (augmentation de la discrimination moyenne des questions) avait également comme motivation de créer de meilleures questions. Il avait un souci marqué pour la création de questions permettant de bien évaluer les étudiants et donc de former de meilleurs médecins. Ce participant avait conscience de l'impact des évaluations sur l'apprentissage des étudiants. L'utilisation active des consignes de rédaction semble moduler la tâche de rédaction de questions chez nos participants. Ainsi, ils le font de manière plus consciente et avec un objectif plus précis en tête. Ils ne veulent pas tout simplement respecter les consignes ou construire des questions claires selon leur point de vue, ils sont motivés par un désir intrinsèque de faire de bonnes évaluations. Ils ont une réflexion plus approfondie quant au processus de création de questions d'examen écrits.

Le participant chez qui l'utilisation de consignes de rédaction a eu le moins d'impact est le participant 3, c'est-à-dire que la moyenne des coefficients de discrimination a été la moins changée de tous les participants. Ce participant démontrait le plus d'hésitation à une adhésion – sans questionnement et dans tous contextes – des consignes de rédaction. Les

propos du participant 3 laissent à croire qu'il consentait à respecter les consignes de rédaction lorsque celles-ci respectaient ses connaissances antérieures et sa vision de la rédaction de questions. Il préférait ne pas respecter une consigne plutôt qu'aller à l'encontre de son jugement. Cependant, pour l'ensemble des questions rédigées par ce rédacteur, un bon respect des consignes a été observé, malgré certaines réticences. On pourrait croire qu'il respectait naturellement par le passé la majorité des consignes présentes dans le guide. Toutefois, ce participant suggère qu'une utilisation trop systématique des consignes pourrait générer des questions identiques qui tendent à évaluer une seule chose ou d'un seul angle.

Les résultats démontrent également que pour le participant 1, la diminution de la discrimination moyenne des questions est moins importante pour l'année 2015-2016. Pour ce participant, une motivation plutôt extrinsèque semblait le pousser à utiliser les consignes de rédaction, c'est-à-dire qu'il avait la particularité de vouloir absolument respecter toutes les règles émises par le programme de médecine par rapport à la rédaction de questions d'examens écrits. L'étude lui a permis d'avoir sous les yeux de manière plus concrète et plus évidente ce qui est recommandé de suivre pour rédiger de bonnes questions. Ainsi, le fait de participer à l'étude, et d'utiliser les consignes, a peut-être permis au participant 1 de rédiger de meilleures questions d'examens sans qu'il s'en rendre nécessairement compte, car elles respectaient selon lui les règles reliées à cette tâche.

5. Conclusion

L'utilisation de consignes lors de la rédaction de questions d'examens écrits semble avoir un impact sur les propriétés psychométriques de celles-ci, et cet impact semble être modulé par l'utilisation (active versus passive) qui est faite des consignes. Une utilisation active amènerait vraisemblablement une augmentation de la discrimination moyenne des questions tandis qu'une utilisation plus passive des consignes de rédaction pourrait diminuer la discrimination moyenne des questions.

Une des limites de cette étude est qu'il est très difficile de contrôler les habitudes de rédaction fortement ancrées chez les rédacteurs. Cependant, les entrevues individuelles ont permis d'explorer ce que la rédaction de questions représentait pour eux et de faire des parallèles avec les analyses quantitatives. Une autre limite qui pourrait être reprochée à cette étude est qu'il n'était pas possible de vérifier l'équivalence des différentes cohortes, soit les étudiants ayant répondu aux questions pour l'année 2014-2015 et ceux pour l'année suivante ce qui aurait pu avoir un impact sur les discriminations de chacune de ces années. Toutefois, les moyennes des examens étaient semblables d'une année à l'autre, ce qui pourrait laisser croire en une similitude entre les cohortes. De plus, lors de l'analyse des questions en fonction du respect des consignes, les trois dernières consignes présentes dans le guide pour les QCMs n'ont pas été évaluées, car elles font référence au contenu de la question et que l'auteure (EVL) n'est pas une experte de contenu. Finalement, nous avons recruté quatre des 13 rédacteurs potentiels pour le programme de médecine ; or on pourrait stipuler un biais positif envers l'utilisation des consignes et un biais de désirabilité sociale. Toutefois, nous avons remarqué dans nos résultats que deux, des quatre participants,

parlaient ouvertement de leurs hésitations, et par conséquent, leur utilisation plutôt passive des consignes de rédaction présentées.

Les résultats obtenus offrent ainsi une nouvelle alternative au monitorage de la qualité des questions *a posteriori* leur administration. Les rédacteurs de questions se voient mieux outillés, et lorsqu'ils font une utilisation active des consignes de rédaction il semble y avoir un potentiel pour un impact positif sur la qualité de leurs questions. Ainsi avec les consignes de rédaction, les rédacteurs se voient maintenant mieux outillés pour que leurs questions soient de meilleure qualité. Comme observé, c'est l'utilisation que chaque rédacteur fait des consignes de rédaction plus que du simple respect des celles-ci qui influence la qualité des questions, il faudrait investiguer davantage sur cet aspect de la rédaction de questions. Une des pistes de recherche futures se trouve dans la façon dont les consignes seront présentées aux nouveaux et plus expérimentés rédacteurs, mais surtout comment les rédacteurs seront amenés à les utiliser pour maximiser leur potentiel.

Références

Bertrand, R. et Blais, J-G. *Modèles de Mesure: L'Apport de la Théorie des Réponses aux Items.* Québec : Presse de l'Université du Québec, 2004

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, *5*(4), 238–242. https://doi.org/10.1016/j.nepr.2004.12.005

Case, S. M., & Swanson, D. B. (2002). Constructing Written Test Questions For the Basic and Clinical Sciences. *Director*, *27*(21), 1–181. Retrieved from http://www.nbme.org/PDF/ItemWriting_2003/2003IWGwhole.pdf

Committee on Accreditation of Canadian Medical Schools (CACMS). (2015) *CACMS Standards and Elements Standards for Accreditation of Medical Education Programs Leading to the M.D. Degree.* https://afmc.ca/pdf/CACMS_Standards_and_Elements_June_2014_Effective_July12015.pdf

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine : Journal of the Association of American Medical Colleges*, *77*(10 Suppl), S103–S104. https://doi.org/10.1097/00001888-200210001-00032

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

Epstein, R. M., & Hundert, E. M. (2002). Professional Competence. *JAMA : The Journal of the American Medical Association*, *287*(2), 226–235.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Laveault, D. & Grégoire, J. (2014). *Introduction aux théories des tests: En psychologie et en sciences de l'éducation.* De Boeck Supérieur

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and Assessment in Teaching*. Merrill. Retrieved from https://books.google.com/books?id=TOZJ3ulKdi8C&pgis=1

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. De Boeck Supérieur. Retrieved from https://books.google.com/books?id=pE6s7zzOyK8C&pgis=1

Vachon Lachiver, E., St-Onge, C., Cloutier, J. et Farand, P. (Soumis 2016). L'identification de consignes discriminantes pour la rédaction d'examens écrits. *Pédagogie médicale*

van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *BMJ (Clinical Research Ed.)*, *321*(7270), 1217–1219. https://doi.org/10.1136/bmj.321.7270.1217

Yin, R. K. (2009). Case study research: Design and methods 4th ed. In *United States: Library of Congress Cataloguing-in-Publication Data*.

ANNEXE 1 : Consignes de rédaction de questions d'examens écrits

Consignes pour la rédaction de Questions à réponse ouverte (QRO)		*Satisfait à l'exigence*
1	Utiliser un vocabulaire approprié pour le niveau du répondant.	
2	Ne pas utiliser de formulation négative.	
3	Rédiger la question de telle sorte qu'il n'existe qu'une seule bonne réponse.	
4	Éviter les réponses qui dépassent une courte phrase.	
5	Indiquer le degré de précision attendu, lorsque pertinent.	
6	Indiquer si l'insertion d'éléments non pertinents sera pénalisée.	

Consignes pour la rédaction de Questions à choix multiples (QCM)		*Satisfait à l'exigence*
1	Les directives dans l'énoncé de la question sont claires et le répondant comprend exactement ce qui est demandé.	
2	Éviter les choix de réponse pouvant susciter du *"test-wiseness"*. Par exemple, éviter les choix de réponse absurdes, invitants (*"formal prompts"*) ou les indices sémantiques (trop spécifiques/généraux).	
3	« Je ne sais pas », « Aucune de ces réponses » ou « Toutes ces réponses » ne doivent pas être un choix de réponse.	
4	Les choix de réponse doivent être mutuellement indépendants/exclusifs.	
5	Éviter de pister la bonne réponse par des constructions grammaticales erronées.	
6	Les choix de réponse doivent inclure seulement une réponse correcte.	
7	Éviter de développer les questions nécessitant des connaissances trop spécifiques.	
8	Éviter les questions créées à partir d'opinion.	

ANNEXE 2 : FORMULAIRE DE CONSENTEMENT

Invitation à participer au projet de recherche

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci.

Chercheuse : Élise Vachon Lachiver, B.Sc.
Étudiante à la Maîtrise en Sciences Cliniques, pédagogie des sciences de la santé
Récipiendaire d'une bourse de la Chaire de Recherche en pédagogie médicale Paul Grand'Maison de la Société des Médecins de l'Université de Sherbrooke.
Faculté de Médecine et des Sciences de la Santé

Équipe de direction : Christina St-Onge, Ph.D.

Madame,
Monsieur,

La qualité des questions d'examens est essentielle pour que l'évaluation des apprentissages et des compétences soit fidèle et valide. Or, il n'est pas toujours facile de rédiger des questions qui ont les qualités psychométriques recherchées. Il appert donc nécessaire d'outiller les personnes responsables de l'évaluation des apprentissages et des compétences à rédiger des questions qui répondent aux règles de l'art, et qui permettent une évaluation qui est fidèle et valide. Nous vous invitons à participer au projet de recherche en titre. Il vise à évaluer l'impact de l'utilisation d'un outil de rédaction de questions d'examens écrits sur les propriétés psychométriques de celles-ci.

Étant responsable d'un examen du programme MD, vous devrez rédiger de nouvelles questions pour cet examen. Votre participation à ce projet de recherche consiste à accepter d'écrire ces questions en utilisant un outil de rédaction de questions que nous vous fournirons. Ce projet est fait dans le cadre de ma maîtrise en Sciences Cliniques sous la supervision de la Pre Christina St-Onge, titulaire de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke. Une rencontre avec la chercheuse principale est prévue afin de vous expliquer l'utilisation de l'outil et de répondre à vos éventuelles questions. Vous devrez, par la suite, faire parvenir une copie de vos questions à la secrétaire du CPSS. Dans un deuxième temps, une entrevue individuelle se fera avec la chercheuse principale afin de percevoir l'utilité et l'acceptabilité de l'outil qui vous a servi à rédiger vos questions d'examens. Finalement, vous consentez à nous donner accès aux coefficients de discrimination des questions qui auront été répondues par les étudiants.

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.

Soyez assuré que votre participation restera confidentielle. Vos questions seront rendues anonymes par la secrétaire (un code sera attribué rendant votre identification impossible). Vos réponses seront incluses dans des résultats de groupe et non pas individuellement. Seuls la chercheuse principale et le personnel de soutien de la Chaire de recherche en pédagogie médicale de la Société des médecins de l'Université de Sherbrooke, tous tenus à la confidentialité, auront accès aux données qui seront conservées sous clé pendant cinq ans après publication. Par la suite, les données seront détruites.

Il n'y a aucun risque connu lié à la participation à ce projet outre le temps que vous y consacrez. De plus, vous pourrez bénéficier de l'utilisation d'un guide de rédaction pour écrire vos questions.

Les résultats de l'étude pourront être communiqués lors de présentations scientifiques (congrès, colloques ou articles).

Si vous désirez plus d'informations sur ce projet, vous pouvez communiquer avec :

Mme Élise Vachon Lachiver, responsable du projet
Étudiante à la Maîtrise en Sciences Clinques, pédagogie des sciences de la santé
Faculté de Médecine et des Sciences de la Santé, FMSS
Université de Sherbrooke

J'ai lu et compris le document d'information au sujet du projet
« L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci ».
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 de participer à l'entrevue individuelle.*

Nom de la participante ou du participant

Signature

Signature de la chercheuse principale

Date

Date

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

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.

Cinquième chapitre : DISCUSSION

Ce projet de maîtrise portait sur la qualité des questions d'examens écrits et incluait deux études complémentaires. Les résultats de ces deux études vous ont d'ailleurs été présentés dans le chapitre précédent. Ce dernier chapitre se veut donc une synthèse des différents objectifs, en plus d'aborder les forces et limites de ce projet. Finalement, les perspectives futures, de même que les retombées pratiques seront présentées.

5.1. Retour sur les objectifs

Les objectifs de cette étude étaient

1) d'identifier les consignes de rédaction de questions d'examens écrits qui permettent de discriminer entre les questions de haute et de faible qualité,

2) de qualifier l'utilisation et l'acceptabilité auprès des rédacteurs des consignes qui ont été préalablement identifiées comme discriminantes, et

3) de quantifier l'impact de leur utilisation sur les propriétés psychométriques des questions lors de la rédaction d'examens écrits.

5.2. Synthèse des conclusions

Les résultats de la première étude nous laissent penser que certaines consignes de rédaction de questions d'examens peuvent discriminer entre les bonnes et moins bonnes questions, ce qui pourrait aider à l'élaboration d'examens de plus haute qualité. Plus spécifiquement, nous avons retenu 14 consignes de rédaction de questions — dont huit se rattachant aux QCM et six pour les QROC – pour l'étude 2.

Lors de la deuxième étude, nous avons remarqué un bon niveau de respect des consignes par tous les participants. La plus grande différence observée entre les participants se trouve au niveau des motifs et motivations qui poussaient les rédacteurs à utiliser les consignes lors de la rédaction de questions ainsi que l'utilité perçue des consignes. Plus spécifiquement, nous avons observé deux types d'utilisateurs de rédaction de consignes. Qui plus est, il semblait y avoir une tendance suggérant que le type d'utilisation des

consignes peut avoir un impact sur la qualité des propriétés psychométriques des questions. Plus spécifiquement, nous avons observé une augmentation du coefficient de discrimination (quoique non significative) chez les rédacteurs qui avaient fait une utilisation active des consignes de rédaction. De plus, nous avons observé une diminution du coefficient de discrimination (quoique non significative) pour les rédacteurs qui en avaient fait une utilisation plus passive des consignes de rédaction. Ainsi, l'impact de l'utilisation des consignes sur la qualité des questions semble être influencé par la manière dont le rédacteur s'en sert.

5.3. Forces et limites

Il s'agit tout d'abord de la première étude du genre à avoir été faite. En effet, l'étude de l'utilisation de consignes de rédaction discriminantes au moment même où les rédacteurs écrivent leurs nouvelles questions dans le but d'en vérifier l'impact sur la qualité des celles-ci est quelque chose de nouveau dans la littérature scientifique tant en pédagogie des sciences de la santé qu'en d'autres domaines. Il s'agit d'une nouvelle avenue pour aider les rédacteurs tout en intervenant en première ligne sur la qualité des questions et donc sur les évaluations. En plus, nous avons été en mesure d'identifier 14 consignes de rédaction de questions d'examens écrits, parmi plus d'une centaine de consignes, qui sont plus discriminantes. De surcroît, cette identification a été faite au moyen d'un processus rigoureux, réalisé en plusieurs étapes. Ceci a permis d'en faire un outil pratique et facile d'utilisation pour les rédacteurs de questions. Une autre force de l'étude est d'avoir obtenu le point de vue des rédacteurs à propos des consignes de rédaction de questions qu'on leur a présentées. En effet, selon van der Vleuten (1996) l'acceptabilité et la faisabilité sont des facteurs importants dans l'adoption d'outils pédagogiques. En ce sens, nous avons recruté quatre de 13 rédacteurs potentiels pour le programme de médecine. Or, on pourrait envisager l'existence d'un biais positif envers l'utilisation des consignes et un biais de désirabilité sociale. Toutefois, nous avons remarqué dans nos résultats que deux (des quatre participants) s'étaient ouverts quant à leurs hésitations, et par conséquent, leur utilisation plutôt passive des consignes de rédaction présentées. Également, la participation de quatre rédacteurs nous assure la réplicabilité des études de cas (Yin, 2009).

Une des limites de l'étude est que les questions d'examens qui ont été rédigées étaient à la base très bonnes, c'est-à-dire qu'elles respectaient déjà un bon nombre de consignes de rédaction. Il y a ainsi eu un certain effet de plafond par rapport au respect (ou non-respect) des consignes de rédaction ce qui pourrait avoir atténué les changements des indices de discrimination observés. Les processus rigoureux de la recherche ainsi que ceux du programme de médecine de l'Université de Sherbrooke ont certainement contribué à la bonne qualité des questions d'examens écrits, ce qui en soi n'est pas une si mauvaise chose étant donné que ces questions font réellement partie des évaluations des étudiants. Une autre limite qui pourrait être reprochée à cette étude est qu'il est très difficile de contrôler les habitudes de rédaction fortement ancrées chez les rédacteurs. Cependant, les entrevues individuelles ont permis d'explorer ce que la rédaction de questions représentait pour eux et de faire des parallèles avec les analyses quantitatives.

5.4. Perspectives futures

Il serait intéressant dans les étapes futures de vérifier les tendances observées dans cette étude sur un plus grand échantillon de rédacteurs ou encore de fournir les consignes de rédaction de questions d'examens écrits à des rédacteurs de questions d'un autre programme en sciences de la santé, comme les soins infirmiers ou les programmes de réadaptation pour ne nommer que ceux-là.

Comme les guides de consignes semblent être perçues positivement par les rédacteurs d'expérience, il serait également pertinent de les offrir à des rédacteurs de questions plus novices. Une comparaison de la qualité des questions pourrait ainsi être faite afin de déterminer si la distribution de ces consignes à tous les rédacteurs de questions est plus bénéfique pour les jeunes rédacteurs qu'à ceux d'expérience.

Finalement, comme nous avons observé que c'est l'utilisation que chaque rédacteur fait des consignes de rédaction plus que du simple respect des consignes qui influence la qualité des questions, il faudrait investiguer davantage sur cet aspect de la rédaction de questions.

5.5. Retombées

La plus importante retombée de ce projet est la création d'un outil pratique pour les rédacteurs de questions. Les consignes de rédaction de questions, présentées comme elles l'ont été dans cette étude, permettent une utilisation facile, simple et rassurante des connaissances concernant la rédaction de questions d'examens écrits de type QROC et QCM. Plusieurs rédacteurs peuvent se sentir dépasser par cette tâche et ainsi rédiger des questions de moins bonne qualité puisqu'ils ne savent pas vraiment comment s'y prendre. Or, les consignes identifiées comme les plus discriminantes offrent un bon coup de pouce à tous les rédacteurs de questions d'examens écrits, qu'ils aient peu ou beaucoup d'expérience. D'ailleurs, la distribution de l'outil auprès des rédacteurs du programme de médecine de l'Université de Sherbrooke est présentement en cours de développement. Ainsi, une feuille synthèse pourra être remise à tous les rédacteurs qui souhaitent l'obtenir pour la rédaction de questions d'examens écrits. La possibilité d'offrir les guides de consignes à d'autres programmes en sciences de la santé est également quelque chose qui est envisagée.

Sixième chapitre : CONCLUSION

En terminant, nous avons pu identifier quatorze consignes de rédaction discriminante de questions de haute et de faible qualité. Par le fait même, des données probantes (14 consignes de rédaction discriminantes) sont maintenant proposées aux rédacteurs sur lesquelles ils peuvent se baser pour rédiger leurs nouvelles questions d'examens écrits. De plus, l'utilisation de consignes de rédaction de questions tend à influencer différemment la qualité des questions écrites selon la manière dont elles ont été utilisées. Ainsi, une utilisation plus active ou à visée plus pédagogique aurait tendance à améliorer la qualité des questions d'un point de vue des propriétés psychométriques selon les résultats obtenus avec nos quatre rédacteurs. Avec la présentation des quatorze consignes, les rédacteurs ont maintenant un outil de plus afin d'aspirer avoir un impact plus positif sur la qualité des questions. Il reste néanmoins du travail à faire afin de confirmer les tendances qui se dégagent de ce mémoire sur un plus grand nombre de rédacteurs de questions d'examens écrits. De plus, la manière dont on présentera ces consignes de rédaction et comment on amènera les rédacteurs à les utiliser à leur plus grand potentiel sont d'autres avenues intéressantes qui restent à investiguer.

LISTE DES RÉFÉRENCES

Amin, Z., Hoon Eng, K., Gwee, M., Chay Hoon, T., & Dow Rhoon, K. (2006). Addressing the needs and priorities of medical teachers through a collaborative intensive faculty development programme. *Medical teacher*, *28*(1), 85-88.

Auger, R. *Formation de base en évaluation des apprentissages: bref historique de l'évaluation des apprentissages, module I*, Les Éditions Logiques, Montréal, 2000.

Bertrand, R. et Blais, J-G. *Modèles de Mesure: L'Apport de la Théorie des Réponses aux Items.* Québec : Presse de l'Université du Québec, 2004

Brady, A. M. (2005). Assessment of learning with multiple-choice questions. *Nurse Education in Practice*, *5*(4), 238–242. https://doi.org/10.1016/j.nepr.2004.12.005

Bush, M. E. (2006). Quality assurance of multiple-choice tests. *Quality Assurance in Education*, *14*(4), 398–404. https://doi.org/10.1108/09684880610703974

Caldwell, D. J., & Pate, A. N. (2013). Effects of question formats on student and item performance. *American Journal of Pharmaceutical Education*, *77*(4), 1–5. https://doi.org/10.5688/ajpe77471

Case, S., & Swanson, D. (1998). Constructing written test questions for the basic and clinical sciences. Retrieved from http://ibmi3.mf.uni-lj.si/mf/fakulteta/prenova/stomatologija/mcq.pdf

Committee on Accreditation of Canadian Medical Schools (CACMS). (2015) *CACMS Standards and Elements Standards for Accreditation of Medical Education Programs Leading to the M.D. Degree.* https://afmc.ca/pdf/CACMS_Standards_and_Elements_June_2014_Effective_July12015.pdf

Conseil de recherches en sciences humaines du Canada, Conseil de recherches en sciences naturelles et en génie du Canada, Instituts de recherche en santé du Canada : *Énoncé de politique des trois Conseils : Éthique de la recherche avec des êtres humains,* décembre 2014.

Conseil médical du Canada. *Lignes directrices pour l'élaboration de questions à choix de réponse,* Février 2010

Considine, J., Botti, M., & Thomas, S. (2005). Design, format, validity and reliability of multiple choice questions for use in nursing research and education. *Collegian*, *12*(1), 19–24. https://doi.org/10.1016/S1322-7696(08)60478-3

Crocker, L., & Algina, J. (1986). Introduction to classical and modern test theory. *Introduction to Classical and Modern Test Theory*. Retrieved from http://ezproxy.usherbrooke.ca/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED312281&lang=fr&site=ehost-live&scope=site

Delgado, A. R., & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment*, *14*(3), 197–201. https://doi.org/10.1027/1015-5759.14.3.197

DiBattista, D., Sinnige-Egger, J.-A., & Fortuna, G. (2014). The "None of the Above" Option in Multiple-Choice Testing: An Experimental Study. *The Journal of Experimental Education*, *82*(2), 168–183. https://doi.org/10.1080/00220973.2013.795127

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do

multiple-choice item-writing principles make any difference? *Academic Medicine : Journal of the Association of American Medical Colleges*, *77*(10 Suppl), S103–S104. https://doi.org/10.1097/00001888-200210001-00032

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

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143. https://doi.org/10.1007/s10459-004-4019-5

Ebel, R. L. (86). Frisbie, DA (1986). *Essential of Educational Measurement*.

Epstein, R. M., & Hundert, E. M. (2002). Professional Competence. *JAMA : The Journal of the American Medical Association*, *287*(2), 226–235.

Gierl, M.J., Zhou, J., Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *Journal of Technology, Learning, and Assessment,* 7(2). Retrieved [date] from http://www.jtla.org

Gierl, M. J., & Lai, H. (2013). Evaluating the quality of medical multiple- choice items created with automated processes. *Medical education*, 47(7), 726-733

Haladyna, T. M., & Downing, S. M. (1989). A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2(1), 37-50.

Haladyna, T. M., & Downing, S. M. (1989). Validity of a Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*. https://doi.org/10.1207/s15324818ame0201_4

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*(3), 309–333. https://doi.org/10.1207/S15324818AME1503_5

Jouquan, J. (2002). L'évaluation des apprentissages des étudiants en formation médicale initiale. *Pédagogie Médicale*, *3*(1), 38–52. https://doi.org/10.1051/pmed:2002006

Laveault, D. & Grégoire, J. (2014). *Introduction aux théories des tests: En psychologie et en sciences de l'éducation.* De Boeck Supérieur

Linn, R. L., & Gronlund, N. E. (2000). *Measurement and Assessment in Teaching*. Merrill. Retrieved from https://books.google.com/books?id=TOZJ3ulKdi8C&pgis=1

Luecht, R.M. (2009). Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm. *In D. J. Weiss (Ed.),* Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

McConnell, M. M., St-Onge, C., & Young, M. E. (2014). The benefits of testing for learning on later performance. *Advances in Health Sciences Education*, 1-16. Doi: 10.1007/s10459-014-9529-1

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. Academic Medicine, 65(9), S63-7.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New Guidelines for Developing Multiple-Choice\nItems. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *2*(2), 65–72. https://doi.org/10.1027/1614-2241.2.2.65

National Board of Medical Examiners (NBME). (2003). Constructing Written Test Questions For the Basic and Clinical Sciences. Third Edition.

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. *Med Teach*, *33*(3), 206–214. https://doi.org/10.3109/0142159X.2011.551559

Prihoda, T. J., Pinckard, R. N., McMahan, C. A., & Jones, A. C. (2006). Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. *Journal of Dental Education*, *70*(4), 378–386.

Québec. Ministère de l'Éducation. (2003) *Politique d'évaluation des apprentissage.* http://feepeq.com/wp/wp-content/uploads/2016/01/PolitiqueEvaluationApprentissages.pdf

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice*, *24*(2), 3–13. https://doi.org/10.1111/j.1745-3992.2005.00006.x

Scallon, G. (2004). *L'évaluation des apprentissages dans une approche par compétences*. De Boeck Supérieur.

Scallon, G. (2000). *L'évaluation formative. Pratiques pédagogiques.* De Boeck Supérieur.

Schneid, S. D., Armour, C., Park, Y. S., Yudkowsky, R., & Bordage, G. (2014). Reducing the number of options on multiple-choice questions: response time, psychometrics and standard setting. *Medical Education*, *48*(10), 1020–1027. https://doi.org/10.1111/medu.12525

Scriven, M. (1967). *The methodology of evaluation. R.W. Tyler, R M. Gagne, M. Scriven (eds.)*, Perspectives of curriculum evaluation, pp.39-83. Chicago, IL: Rand McNally.

St-Onge, C., Cloutier, J., & Côté, D. J. (2013). La formation des rédacteurs peut-elle aider à améliorer la rédaction de questions à réponse écrite? Présentation orale faite au 81ème congrès de l'ACFAS, Québec, Québec. (2013-05-27)

Tardif, J. (2006). L'évaluation des compétences. *Documenter le parcours de développement.* Montréal: Chenelière Éducation.

Tarrant, M., Ware, J., & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, *9*, 40. https://doi.org/10.1186/1472-6920-9-40

Vachon Lachiver, E., St-Onge, C., Cloutier, J. et Farand, P. (Soumis 2016). L'identification de consignes discriminantes pour la rédaction d'examens écrits. *Pédagogie médicale*

van der Vleuten, C. (2000). Validity of final examinations in undergraduate medical training. *BMJ (Clinical Research Ed.)*, *321*(7270), 1217–1219. https://doi.org/10.1136/bmj.321.7270.1217

Wood, T. (2009). Assessment not only drives learning, it may also help learning. *Medical Education*, 43(1), 5-6.

Yin, R. K. (2009). Case study research: Design and methods 4th ed. In *United States: Library of Congress Cataloguing-in-Publication Data*.

ANNEXES

ANNEXE 1 : CERTIFICAT D'APPROBATION ÉTHIQUE

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 :

L'utilisation de critères de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci.

Élise Vachon Lachiver
Étudiante à la maîtrise 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
Mathieu Gagnon, professeur à la Faculté d'éducation, Département d'enseignement au préscolaire et au primaire
Suzanne Guillemette, professeure à la Faculté d'éducation, Département de gestion de l'éducation et de la formation
Sawsen Lakhal, professeure à la Faculté d'éducation, Département de pédagogie
Mélanie Lapalme, professeure à la Faculté d'éducation, Département de psychoéducation
Julie Myre-Bisaillon, 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, 2 juillet 2015

ANNEXE 2 : FORMULAIRE DE CONSENTEMENT

Invitation à participer au projet de recherche et formulaire de consentement

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci.

Chercheuse : Élise Vachon Lachiver, B.Sc.
Étudiante à la Maîtrise en Sciences Cliniques, pédagogie des sciences de la santé
Récipiendaire d'une bourse de la Chaire de Recherche en pédagogie médicale Paul Grand'Maison
de la Société des Médecins de l'Université de Sherbrooke.
Faculté de Médecine et des Sciences de la Santé

Équipe de direction : Christina St-Onge, Ph.D.

Madame,
Monsieur,

La qualité des questions d'examens est essentielle pour que l'évaluation des apprentissages et des compétences soit fidèle et valide. Or, il n'est pas toujours facile de rédiger des questions qui ont les qualités psychométriques recherchées. Il appert donc nécessaire d'outiller les personnes responsables de l'évaluation des apprentissages et des compétences à rédiger des questions qui répondent aux règles de l'art, et qui permettent une évaluation qui est fidèle et valide. Nous vous invitons à participer au projet de recherche en titre. Il vise à évaluer l'impact de l'utilisation d'un outil de rédaction de questions d'examens écrits sur les propriétés psychométriques de celles-ci.

Étant responsable d'un examen du programme MD, vous devrez rédiger de nouvelles questions pour cet examen. Votre participation à ce projet de recherche consiste à accepter d'écrire ces questions en utilisant un outil de rédaction de questions que nous vous fournirons. Ce projet est fait dans le cadre de ma maîtrise en Sciences Cliniques sous la supervision de la Pre Christina St-Onge, titulaire de la Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des médecins de l'Université de Sherbrooke. Une rencontre avec la chercheuse principale est prévue afin de vous expliquer l'utilisation de l'outil et de répondre à vos éventuelles questions. Vous devrez, par la suite, faire parvenir une copie de vos questions à la secrétaire du CPSS. Dans un deuxième temps, une entrevue individuelle se fera avec la chercheuse principale afin de percevoir l'utilité et l'acceptabilité de l'outil qui vous a servi à rédiger vos questions d'examens. Finalement, vous consentez à nous donner accès aux coefficients de discrimination des questions qui auront été répondues par les étudiants.

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.

Soyez assuré que votre participation restera confidentielle. Vos questions seront rendues anonymes par la secrétaire (un code sera attribué rendant votre identification impossible). Vos réponses seront incluses dans des résultats de groupe et non pas individuellement. Seuls la chercheuse principale et le personnel de soutien de la Chaire de recherche en pédagogie médicale de la Société des médecins de l'Université de Sherbrooke, tous tenus à la confidentialité, auront accès aux données qui seront conservées sous clé pendant cinq ans après publication. Par la suite, les données seront détruites.

Il n'y a aucun risque connu lié à la participation à ce projet outre le temps que vous y consacrez. De plus, vous pourrez bénéficier de l'utilisation d'un guide de rédaction pour écrire vos questions.

Les résultats de l'étude pourront être communiqués lors de présentations scientifiques (congrès, colloques ou articles).

Si vous désirez plus d'informations sur ce projet, vous pouvez communiquer avec :

Mme Élise Vachon Lachiver, responsable du projet
Étudiante à la Maîtrise en Sciences Clinques, pédagogie des sciences de la santé
Faculté de Médecine et des Sciences de la Santé, FMSS
Université de Sherbrooke

J'ai lu et compris le document d'information au sujet du projet
« L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci ».
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 de participer à l'entrevue individuelle.*

Nom de la participante ou du participant

Signature

Date

Signature de la chercheuse principale

Date

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

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.

ANNEXE 3 : PROTOCOLE PREMIÈRE RENCONTRE

Rédaction de questions d'examens écrits : WEQQ-R

Bienvenue à cette rencontre explicative concernant le projet de recherche sur la rédaction de questions d'examens écrits. Mon nom est Élise Vachon Lachiver et je suis la responsable du projet de recherche ainsi que du déroulement de la rencontre d'aujourd'hui. Cette rencontre durera environ 30 minutes.

Comme vous le savez, l'élaboration de bonnes questions d'examens est une tâche très ardue, mais également un enjeu important pour le programme de médecine. La construction des examens demande beaucoup de temps, d'efforts et de ressources de votre part. Nous désirons ainsi vous aider en vous proposant une grille de rédaction de questions pour faciliter (baliser) cette tâche et contribuer à augmenter la qualité des questions.

Plus spécifiquement, nous désirons réaliser un projet de recherche dont l'objectif est de déterminer si l'utilisation de consignes pendant la rédaction de questions d'examens écrits améliore les propriétés psychométriques des dites questions. Un second objectif vise à qualifier l'utilisation et l'acceptabilité de l'outil auprès des utilisateurs de la grille, c'est-à-dire vous-mêmes, les rédacteurs.

Avant toute chose, j'aimerais préciser quelques points :

- *d'abord, je vous lirai le formulaire de consentement pour vous informer de l'étude et de votre tâche.*
- *de plus, je vous expliquerai comment vous servir de l'outil que vous aurez à utiliser pour la rédaction.*
- *par la suite, vous répondrez à un court questionnaire sur votre expérience en tant que rédacteur de questions d'examens écrits.*
- *finalement, nous prendrons rendez-vous pour l'entrevue individuelle si vous acceptez de faire cette partie de l'étude.*

Signer le formulaire de consentement

Voici un formulaire de consentement concernant ce projet de recherche qui en lien avec la tâche que vous devrez effectuer. Nous allons le lire attentivement ensemble. Si vous avez des questions, n'hésitez pas à me les poser.

Remise d'une copie au participant.

L'assistant de recherche lit lentement et à voix haute le formulaire de consentement.
Accent sur le titre du projet, le nom de la responsable et les coordonnées du comité d'éthique.

Acceptez-vous de faire une entrevue portant sur votre appréciation de l'utilisation du WEQQ-R pendant votre rédaction?

Faire signer les **2 copies** du formulaire en remettre une au participant.

Explication et familiarisation de l'utilisation de l'outil WEQQ-R

Je vous explique maintenant plus précisément comment doit être utilisé l'outil de rédaction de questions, le Written Exam Question Quality-Revised (WEQQ-R). Voici votre copie de l'outil.

Remise d'un exemplaire de l'outil au participant.
L'outil est construit sous forme de grille à cocher pour s'assurer que tous les éléments sont présents pour obtenir une question de qualité. L'utilisation du WEQQ-R se fait en quatre étapes simples.
D'abord, il est recommandé de lire toutes les consignes avant de commencer quoi que ce soit. Vous ferez donc une première lecture des consignes avant d'écrire.
Vous pourrez ensuite rédiger vos nouvelles questions en ayant l'outil à votre portée.
La troisième étape consiste à faire une vérification du respect des consignes de rédaction qui se trouvent dans le WEQQ-R. Vous cochez ainsi toutes les consignes qui sont respectées.
Finalement, vous effectuez des modifications au besoin sur vos questions.

Nous allons maintenant lire chacune de ces consignes de rédaction de questions ensemble afin qu'elles soient bien comprises. Vous pouvez m'arrêter à tout moment si vous avez des questions ou pour avoir plus de précision sur une consigne en particulier. Je vous montrerai par la suite un exemple de question qui respecte les consignes de rédaction ainsi qu'un contre-exemple.

Lecture de toutes les consignes de l'outil.

Comme mentionné précédemment voici maintenant un exemple de question qui respecte toutes les consignes de rédaction qui se trouve dans le WEQQ-R.

Remise de l'exemple au participant. Lecture de celui-ci.

Pour vous faire une idée claire, je vous remets également un contre-exemple de question.

Remise du contre-exemple au participant. Lecture de celui-ci.

Avez-vous des questions?

Questionnaire sur l'expérience en tant que rédacteur de questions d'examens écrits

Maintenant pour obtenir un portrait global des rédacteurs qui participent à notre étude, je vous demanderais de répondre à ce court questionnaire pour mieux connaitre votre expérience de rédaction de questions.

Remettre une copie du questionnaire au participant.

Si vous avez des questions, n'hésitez pas à me les poser.

Prise d'un rendez-vous pour l'entrevue semi-dirigée (pour ceux ayant accepté de faire cette partie de l'étude)

J'aimerais vous rencontrer dans les quatorze jours suivants l'envoi de vos questions au secrétariat du programme de médecine afin que l'on discute de l'utilité et de l'acceptabilité de l'outil qui vous a servi à rédiger vos questions d'examens. Cette rencontre devrait durer 30 min. Avez-vous un moment dans votre horaire qui pourrait convenir à cette rencontre?

Un courriel de rappel indiquant la date, l'heure et l'endroit convenus vous sera envoyé trois jours avant la tenue de l'entrevue individuelle.

Je vous remercie de votre participation à ce projet de recherche.

ANNEXE 4 : EXEMPLES DE QUESTIONS PRÉSENTÉES AUX RÉDACTEURS

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci.

Exemples de Questions à Réponse Ouverte (QRO)

Question répondant aux consignes présentées dans le WEQQ-R

Une petite fille de 6 ans se présente pour un développement mammaire au stade III deTanner et des poils pubiens aussi au stade III. L'investigation sanguine démontre un FSH et un LH bas (niveau prébubaire).

Quel est le type de puberté précoce?

Question **ne répondant pas** aux consignes présentées dans le WEQQ-R

Nommez 2 comportements sexuels inappropriés qui peuvent être présents chez une personne ayant subi un TCC et pour chaque comportement, décrivez une intervention que vous pourriez mettre en œuvre auprès du patient ou de sa famille.

1. _____
   _____
   _____
2. _____
   _____
   _____

Exemples de Questions à Choix Multiples (QCM)

Question répondant aux consignes présentées dans le WEQQ-R

Un patient mentionne qu'il voudrait bien perdre du poids pour améliorer sa tolérance à l'effort mais il ne pense pas pouvoir le faire.

Identifiez à quel stade de changement il se situe:

A. Pré-contemplation
B. Contemplation
C. Préparation
D. Action

Question **ne répondant pas** aux consignes présentées dans le WEQQ-R

Conduire une automobile est :

- ☐ une tâche simple, qui fait principalement appel à mes réflexes et mes automatismes
- ☐ une tâche complexe, exigeant le bon fonctionnement de plusieurs capacités spécifiques
- ☐ une tâche complexe, qui fait principalement appel à mes réflexes et mes automatismes
- ☐ une tâche comme bien d'autres, ne demandant aucune habileté spécifique
- ☐ je ne sais pas

ANNEXE 5 : QUESTIONNAIRE EXPÉRIENCE DES RÉDACTEURS

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci

Questionnaire sur votre expérience en tant que rédacteur de questions d'examens écrits

Veuillez répondre à ce questionnaire qui nous permettra de mieux connaitre votre expérience et ainsi approfondir l'analyse de nos résultats. N'hésitez pas à poser des questions si vous avez des doutes sur la signification de certaines questions. Une fois complétée, remettre le questionnaire à la chercheuse principale de l'étude.

1. Depuis combien d'années rédigez-vous des questions d'examens écrits? _____

2. Pour quelle(s) unité(s) rédigez-vous des questions d'examens?

   _____

   _____

3. Quelles ressources (guide de rédaction, etc) utilisez-vous pour rédiger de nouvelles questions?

   _____

   _____

4. Comment décrivez-vous la tâche de rédaction de question (ex : devoir incontournable, défi stimulant, etc)?

   _____

   _____

5. Quelles sont selon-vous les caractéristiques d'une bonne question (autres que celles discutées)?

   _____

   _____

6. Avez-vous déjà eu une formation concernant la rédaction de questions d'examens écrits?

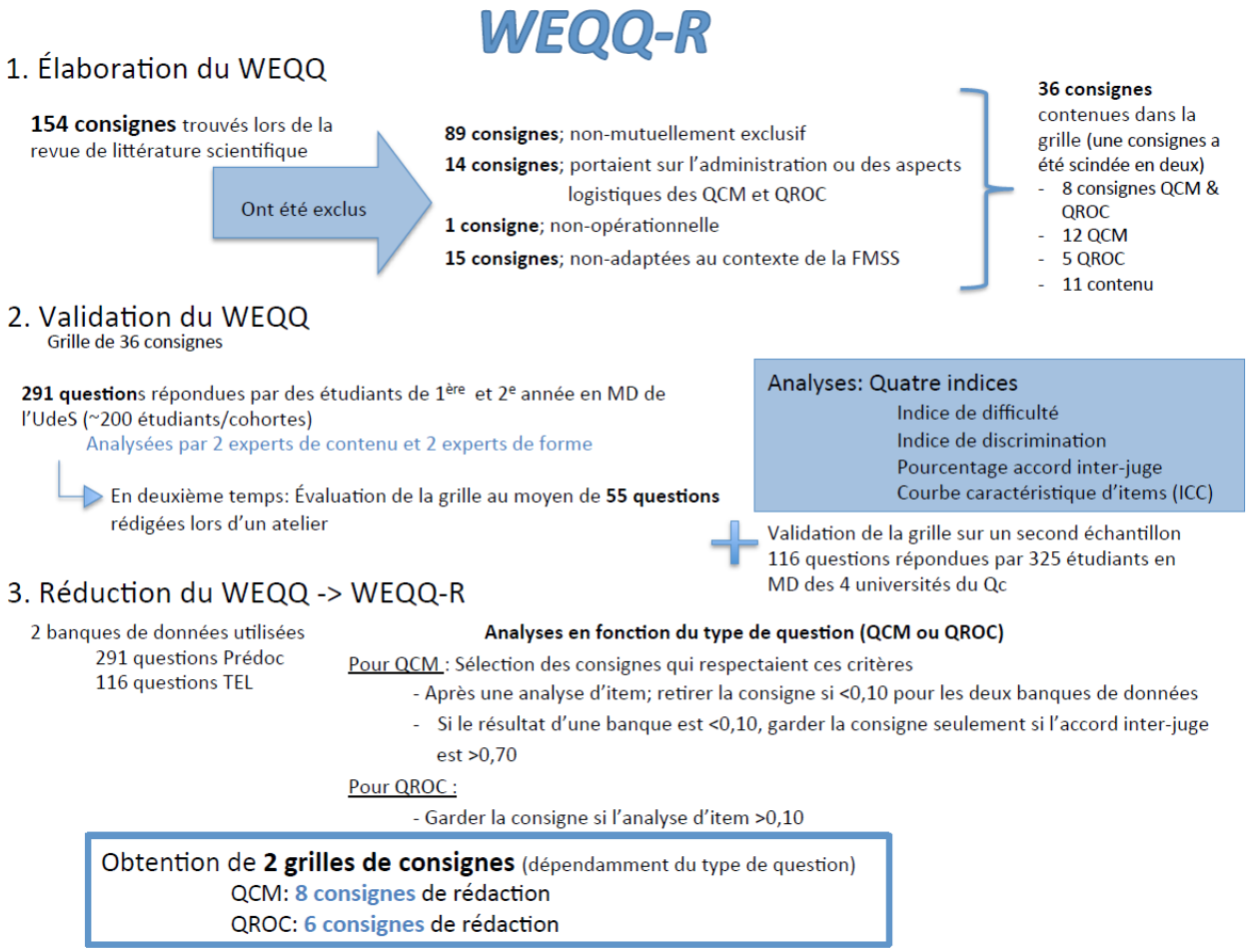
   ❒ Non – Merci d'avoir complété ce questionnaire.

   ❒ Oui – SVP compléter le tableau ci-dessous.

Nom ou contenu de la formation	Durée	Animateur/Organisation

Merci!

ANNEXE 6 : SCHÉMA DE L'OBTENTION DES GRILLES DE CONSIGNES DE RÉDACTION DE QUESTIONS



Étape 3 illustrée ici répond à l'objectif 1 de ce projet de maîtrise et a été réalisée par l'étudiante.

ANNEXE 7 : GRILLES DE CONSIGNES DE RÉDACTION DE QUESTIONS

Ces grilles de consignes ont été obtenues suite à la réalisation de l'objectif 1 de ce projet de maîtrise. Elles ont été élaborées par l'étudiante et sont décrites dans l'article 1 (voir page 20).

Written Exam Question Quality-Revised **(WEQQ-R)**

Grille de consignes pour la qualité des Questions choix multiples (QCM)

	Consignes	Satisfait à l'exigence oui
1	Les directives dans l'énoncé de la question sont claires et le répondant comprend exactement ce qui est demandé.	
2	Éviter les choix de réponse pouvant susciter du "test-wiseness". Par exemple, éviter les choix de réponse absurdes, invitants (formal prompts) ou les indices sémantiques (trop spécifiques/généraux).	
3	"Je ne sais pas", "Aucune de ces réponses" ou "Toutes ces réponses" ne doivent pas être un choix de réponse.	
4	Les choix de réponse doivent être mutuellement indépendants/exclusifs.	
5	Éviter de pister la bonne réponse par des constructions grammaticales erronées.	
6	Les choix de réponse doivent inclure seulement une réponse correcte.	
7	Éviter de développer les questions nécessitant des connaissances trop spécifiques.	
8	Éviter les questions créées à partir d'opinion.	

Vachon Lachiver, E., St-Onge, C. *et al.* (2015), *L'identification de consignes pour faciliter la rédaction d'examens écrits.* [Soumis CCME 2016]

Written Exam Question Quality-Revised **(WEQQ-R)**

Grille de consignes pour la qualité des Questions à Réponse Ouverte (QRO)

	Consignes	Satisfait à l'exigence oui
1	Utiliser un vocabulaire approprié pour le niveau du répondant.	
2	Ne pas utiliser de formulation négative.	
3	Rédiger la question de telle sorte qu'il n'existe qu'une seule bonne réponse.	
4	Éviter les réponses qui dépassent une courte phrase.	
5	Indiquer le degré de précision attendu, lorsque pertinent.	
6	Indiquer si l'insertion d'éléments non pertinents sera pénalisé.	

Vachon Lachiver, E., St-Onge, C. *et al.* (2015), *L'identification de consignes pour faciliter la rédaction d'examens écrits*. [Soumis CCME 2016]

ANNEXE 8 : GUIDE D'ENTREVUE DEUXIÈME RENCONTRE

L'utilisation de consignes de rédaction de questions d'examens écrits en médecine et l'effet sur les propriétés psychométriques de celles-ci

Guide d'entrevue semi-dirigée individuelle

Bienvenue à cette rencontre sous forme d'entrevue individuelle concernant le projet de recherche sur la rédaction de questions d'examens écrits. Mon nom est Élise Vachon Lachiver et je suis la responsable du projet de recherche ainsi que du déroulement de la rencontre d'aujourd'hui. Cette rencontre durera environ 45 minutes.

Cette étude a pour but dans un premier temps de déterminer si l'utilisation de consignes pendant la rédaction de questions d'examens écrits améliore les propriétés psychométriques des dites questions.

Un second objectif vise à qualifier l'utilisation et l'acceptabilité de l'outil, soit du WEQQ-R auprès des rédacteurs. Il s'agit également de l'objectif de la rencontre d'aujourd'hui.

Pour ce faire, je vous poserai des questions auxquelles je vous demande de répondre du mieux que vous pouvez en parlant de manière haute et claire. Cela facilitera la captation audio de cette rencontre. L'enregistrement de l'entrevue se fera uniquement dans un but d'analyse et sera conservée dans un endroit sécurisé.

Avez-vous des questions?
Êtes-vous prêts à commencer?

Démarrer l'enregistreur.

Guide d'entrevue Post-UTILISATION WEQQ-R

A. Oui, le WEQQ-R a été utilisé

1. Comment avez-vous trouvé son utilisation lors de la rédaction de vos questions écrites?
   a. Qu'avez-vous trouvé le plus facile?
   b. Qu'avez-vous trouvé le plus difficile?

2. Qu'est-ce qui vous a motivé à l'utiliser?

3. Pouvez-vous me parler de vos impressions quant aux consignes retenues dans l'outil?
   a. Surprises?
   b. Consignes manquantes?
   c. Non-accord avec certaines consignes?

4. Quel impact, en termes de temps, a eu l'utilisation du WEQQ-R sur votre rédaction de questions d'examens?
   a. Quelles sont les répercussions de cet impact?
   b. Comment jugez-vous cet impact?

5. Vous sentez vous mieux outillé pour rédiger de nouvelles questions maintenant que vous connaissez le WEQQ-R?

6. Quelles améliorations feriez-vous?

7. Qu'avez-vous apprécié de l'outil?
   a. Le plus, le moins?

8. À l'avenir, utiliseriez-vous le WEQQ-R pour rédiger de nouvelles questions?
   a. Pourquoi?

9. Le recommanderiez-vous à vos collègues pour rédiger de nouvelles questions?
   b. Pourquoi?

B. Non, le WEQQ-R n'a pas été utilisé

1. Qu'est-ce qui a fait en sorte que vous n'avez pas utilisé l'outil?

2. Lors de votre prochaine rédaction de questions, pensez-vous utiliser le WEQQ-R?

3. Que diriez-vous à un collège qui se demande s'il doit utiliser le WEQQ-R pour s'aider dans la rédaction de questions d'examens écrits?

4. Avez-vous déjà utilisé d'autres consignes de rédaction?
    a. Quelles étaient ces consignes ou lignes directrices?
    b. Les consignes de rédaction du WEQQ-R entraient-elles en conflit avec celles que vous utilisez habituellement? De quelle manière?

5. Quelles améliorations feriez-vous?

6. Qu'avez-vous apprécié de l'outil?
    c. Le plus, le moins?

QUESTIONS RELATIVES AU PARTICIPANT