Pédagogie Médicale **18**, 55-64 (2017)

https://doi.org/10.1051/pmed/2018002

Pédagogie Médicale

Disponible en ligne :
www.pedagogie-medicale.org

Recherche et perspectives

La rédaction de questions à choix multiple et de questions à réponse ouverte et courte pour les examens écrits dans les programmes de formation en santé : une étude docimologique des consignes de rédaction

Creating multiple-choice questions and short open-answer questions in professional health-education programs: a docimological study of guideline drafting

Élise VACHON LACHIVER[1], Christina ST-ONGE[1,*], Jacinthe CLOUTIER[2], et Paul FARAND[1]

[1] Faculté de médecine et des sciences de la santé, Université de Sherbrooke, Sherbrooke, QC, Canada
[2] Faculté des sciences de l'agriculture et de l'alimentation, Université Laval, Québec, QC, Canada

Manuscrit soumis le 18 août 2016 ; commentaires éditoriaux formulés aux auteurs le 8 mai 2017 et le 4 janvier 2018 ; accepté pour publication le 17 février 2018

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 que l'on retrouve dans la littérature scientifique plus de 150 consignes destinées à en guider 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. **Méthodes :** 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) provenant de banques de questions du programme de médecine de notre institution ainsi que 116 questions QCM provenant d'un précédent projet de recherche. 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 pour les QROC (discrimination moyenne = 0,20) et huit consignes de rédaction pour les QCM (discrimination moyenne = 0,19 ; accord inter-juges moyen = 0,98) ont été identifiées comme discriminantes. **Conclusions :** Nos résultats suggèrent 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.

Mots clés : évaluation, examens écrits, propriétés psychométriques, étudiants en médecine

Abstract. Background/Aim: The assessment of learning is undeniably important in health sciences education. The development of written questions is a real challenge for those drafting them. Scientific literature has more than 150 instructions that guide the drafting of written questions. Unfortunately, there is little evidence on the impact of such guidelines on the quality of questions. The aim of this study is to identify the guidelines that differentiate well- from poorly-drafted questions, and accordingly need to be looked at in the drafting process. **Method:** We selected 36 guidelines among the 150 listed in the literature. We used the guidelines to assess 407 written questions: 150 multiple-choice questions (MCQ) and 141 short-answer questions (SAQ) from the database in our undergraduate medical education program and 116 others questions (MCQ) from a previous research project. We carried out an items analysis based on the principles of the Classical Test Theory to identify the most discriminant guidelines. **Results:** Six guidelines for SAQ (mean discrimination = 0.20) and 8 guidelines for MCQ (mean discrimination = 0.19; mean inter-rater agreement = 0.98) were identified as discriminant. **Conclusion:** Our results suggest that some of the guidelines for written exams can discriminate between well- and poorly-drafted questions, which could help develop more high-quality evaluations.

Keywords: evaluation, written exams, psychometric properties, medical student

*Correspondance et offprints : Christina ST-ONGE, Service de médecine interne / Département de médecine. Faculté de médecine et des sciences de la santé. Université de Sherbrooke 3001, 12e Avenue Nord Sherbrooke (Québec) J1H 5N4
Mailto : christina.st-onge@usherbrooke.ca.

Introduction

L'évaluation des apprentissages et des compétences est une démarche pédagogique centrale dans le cadre des programmes de formation initiale en sciences de la santé, que ce soit durant le cursus préclinique, dans le cadre des stages cliniques du programme pré-gradué (externat) ou dans celui des stages du programme post-gradué (résidence, résidanat ou internat, selon les pays) [1]. Les évaluations sont utilisées, notamment dans le cursus préclinique, pour assurer un premier point de contrôle des apprentissages des étudiants, c'est-à-dire pour vérifier que les étudiants possèdent les connaissances adéquates et les habiletés nécessaires pour intégrer le cursus clinique et ainsi, participer à la prise en charge des soins accordés aux patients [2–5]. Dans ce cas, l'évaluation peut avoir d'importantes conséquences, par exemple, en conditionnant l'accès de l'étudiant à une profession ou à une spécialisation souhaitée [3]. Au regard d'une telle finalité, dans une perspective docimologique de l'évaluation, les examens doivent pouvoir différencier les étudiants ayant mieux réussi l'évaluation de ceux ayant moins bien réussi cette même évaluation, c'est-à-dire qu'ils doivent être discriminants [6]. Pour répondre à de telles exigences, le développement d'examens dans le cadre de la formation médicale doit être fait de manière rigoureuse et appropriée [1], et satisfaire les plus hauts standards de qualité.

L'élaboration de questions d'examens écrits est souvent décrite comme étant un grand défi pour les rédacteurs [7–8] et elle est souvent considérée comme une tâche ardue [9–11]. En effet, la rédaction de questions d'examens écrits demande beaucoup de temps, d'efforts, de ressources et de connaissances [12–13]. Malheureusement, il arrive que des questions d'examens soient mal rédigées (problèmes de contenu, ambiguïtés, faibles propriétés psychométriques, etc.) et doivent être retirées du dispositif d'évaluation lors de l'analyse psychométrique. Cependant, la modification du contenu de l'évaluation *a posteriori* a des conséquences négatives. Cela diminue notamment la validité de l'évaluation et ce, sur plusieurs plans [5,14–16]. Selon Bertrand et Blais [17], la validité d'un dispositif d'évaluation est satisfaite lorsque le « jugement [est] 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 peut ne plus être 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.

L'une des options qui s'offre aux rédacteurs soucieux de rédiger des questions de qualité est de se tourner vers les écrits scientifiques répertoriant des consignes de rédaction des questions d'examens écrits. On entend par examen écrit tout examen administré de manière telle que l'étudiant doit répondre en utilisant un support de type « papier-crayon ». De fait, les examens comportant des questions à choix multiples (QCM) ou des questions à réponse ouverte et courte (QROC) sont des examens écrits. Plus de 150 consignes de rédaction de ce type de questions ont été répertoriées [1,13–16,18] ; il peut dès lors être difficile pour les rédacteurs de se retrouver dans cette littérature, faute d'expertise et de ressources. De plus, il existe peu de données probantes quant à ces consignes, c'est-à-dire que peu de données nous indiquent si le respect ou non-respect de ces consignes a un impact sur les propriétés psychométriques des questions d'examens écrits.

Ainsi, l'objectif de cette étude est d'identifier les consignes de rédaction de questions de type QCM ou QROC qui permettent de discriminer les questions d'examens écrits de haute et de basse qualité, la restriction de l'étude aux questions de type QROC et QCM étant justifiée à nos yeux de façon pragmatique dès lors que ce se sont, de très loin, les formats d'examens écrits les plus largement répandus dans le cadre des programmes de formation en santé dans le contexte nord-américain.

Méthodes

Devis général

Une analyse d'items, basée sur la théorie classique des tests, a été réalisée dans le cadre de cette étude afin d'obtenir les indices de discrimination pour chaque consigne étudiée. Les résultats de l'analyse d'items ont été utilisés pour identifier les consignes de rédaction de questions de type QCM ou QROC, qui permettent de différencier les questions de « bonne qualité » de celles de « moins bonne qualité ». De façon plus spécifique, il s'agit d'identifier les consignes dont le respect, au moment de la rédaction des questions, est associé à des indicateurs de qualités psychométriques considérés comme élevés, dans une perspective docimologique. Dès lors la qualité des questions devrait pouvoir être inférée à partir du nombre total de consignes respectées. Une question bien rédigée, c'est-à-dire de bonne qualité, devrait ainsi respecter le maximum de consignes de rédaction, étant postulé que la qualité d'une question est proportionnelle au nombre de consignes respectée lors de sa rédaction [19–20].

Inventaire méthodique et critique des consignes de rédaction de questions publiées dans la littérature

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* ». Les recherches ont été effectuées à partir des moteurs de recherche disponibles sur les bases de données 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

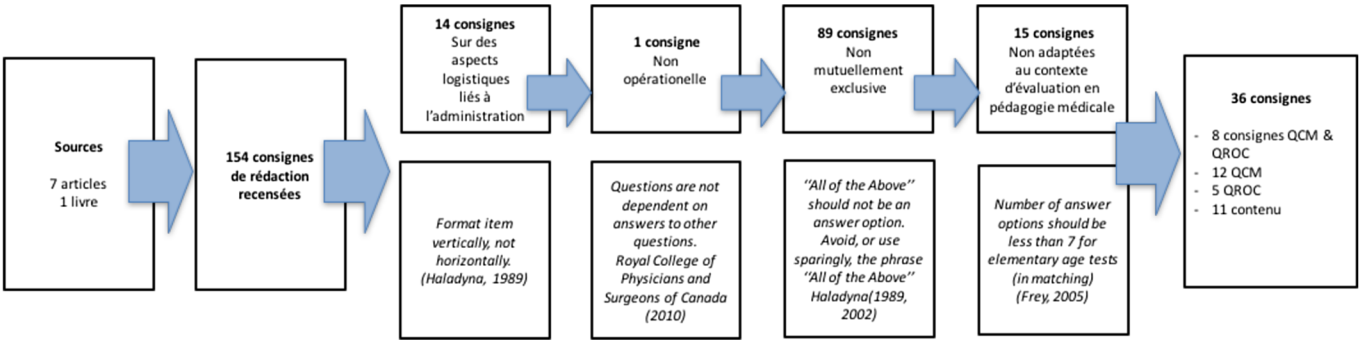

Figure 1. Identification d'un corpus de consignes de rédaction appropriées pour les examens en pédagogie des sciences de la santé. QROC : Question à réponse ouverte courte ; QCM : Question à choix multiples.

Tableau I. 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 précédent (influence du *testing*) [23]
Indices mesurés	Indice de discrimination des consignes de rédaction	Indice de discrimination des consignes de rédaction	Indice de discrimination des consignes de rédaction Accord inter-juges

QROC : Question à réponse ouverte courte ; QCM : Question à choix multiples ; UdeS : Université de Sherbrooke.

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). Rappelons que les données probantes à propos du respect (ou non-respect) des consignes de rédaction des questions sont peu présentes dans la littérature scientifique.

Un processus itératif de sélection des consignes, adapté de la procédure de DeVellis [21], et illustré sur la figure 1, a été développé 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 QCM et des QROC ont été retenues (retrait de 14 consignes de rédaction). Dans un deuxième temps, ces consignes ont été révisées de façon itérative afin de produire une grille d'énoncés opérationnels (consigne claire et applicable ; retrait d'une consigne de rédaction au regard de ce critère), mutuellement exclusifs (évaluant des éléments distincts ; retrait de 89 consignes de rédaction au regard de ce critère) et adaptés au contexte d'évaluation en pédagogie des sciences de la santé (retrait de 15 consignes de rédaction au regard de ce critère). À ce point, les consignes de rédaction liées à la forme et au contenu ont été retenues.

Au terme de ce processus, 35 consignes ont été retenues, transformées en énoncés opérationnels et regroupées en quatre catégories :

– consignes non différenciées (c.-à-d. pouvant s'appliquer aux QCM et QROC) ;
– consignes applicables aux QCM ;
– consignes applicables aux QROC ;
– consignes relatives au contenu (c.-à-d. qui requièrent une expertise de la matière pour être évaluées).

La liste de consignes a été pré-testée par un expert de forme et par un expert de contenu sur une trentaine de questions. Après cette première mise à l'essai sur un petit nombre de questions, quelques correctifs (par exemple : changement de mots) ont été apportés pour assurer une compréhension uniforme des consignes et, par conséquent, favoriser une utilisation standardisée de celles-ci. La liste finale comportait 36 consignes réparties en quatre catégories. Un énoncé de la grille préliminaire a été scindé en deux afin d'en améliorer sa compréhension et faciliter son utilisation. Cet énoncé, tel qu'initialement formulé, avait deux *focus* et pouvait porter à confusion. La catégorie « non différenciée » contenait huit consignes, la catégorie « QCM » en contenait 12, la catégorie « QROC » en contenait cinq et la catégorie « contenu » contenait 11 consignes. Comme l'objectif de cette liste est d'inférer le niveau de qualité d'une question à partir du respect ou non des consignes, une échelle binaire de codage, selon que la question respecte (codé 1) ou non (codé 0) la consigne, a été établie.

Corpus des questions d'examens analysées dans l'étude

Trois banques de questions ont fait office de bases de données dans cette étude. Leurs caractéristiques sont résumées dans le tableau I. Un total de 407 questions différentes a été considéré. Aucune donnée relative au niveau des étudiants ou des groupes concernés n'a été exploitée au

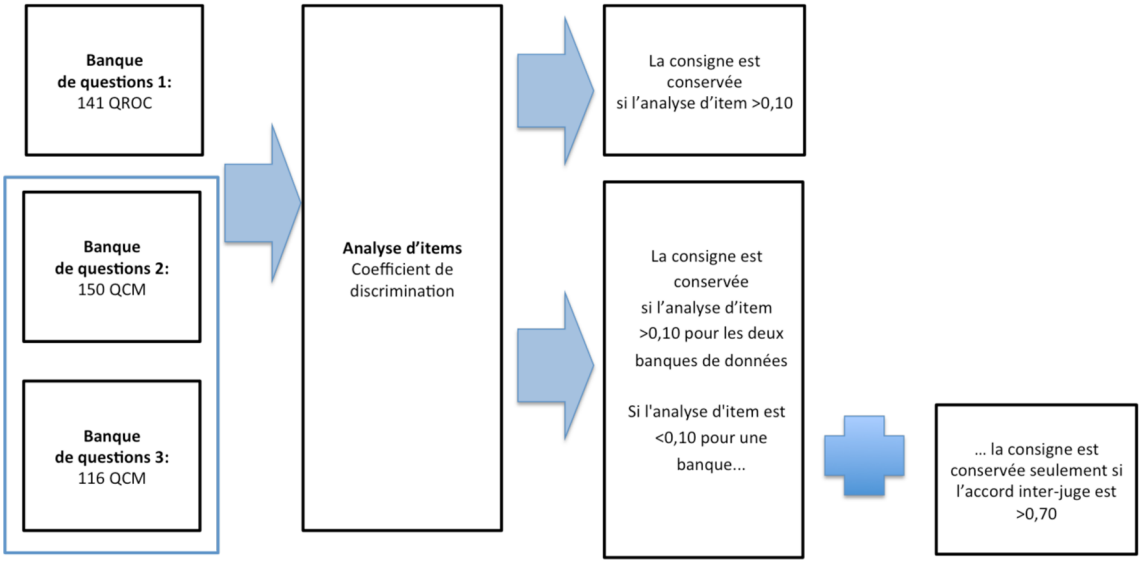

Figure 2. Sélection des consignes les plus discriminantes. QROC : Question à réponse ouverte courte ; QCM : Question à choix multiples.

cours de cette étude ; seules les questions ont été utilisées afin de calculer l'indice de discrimination de chacune des consignes. Ainsi, aucune autorisation d'un comité d'éthique n'a été sollicitée, conformément à la politique d'éthique des trois conseils de recherche du Canada [22].

Banque de questions 1

Cent quarante-et-une 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).

Banque de questions 2

Cent cinquante QCM ont également été choisies aléatoirement dans la même banque de questions du programme préclinique de médecine de l'UdeS. Ces questions provenaient des mêmes unités d'enseignement susmentionnées.

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 qui ont servi dans le cadre d'une étude sur l'influence du *testing* [23].

Procédure

Pour chacune des 407 questions décrites ci-dessus, nous avons déterminé si elles respectaient (ou non) chacune des 36 consignes de rédaction retenues dans chacune des banques de données respectivement. Ces jugements (respectivement cotés 0 et 1) ont été utilisés dans le cadre de l'analyse d'items [17]. Les juges devaient ainsi évaluer si les questions respectaient les consignes présentes dans la liste. Nous avons eu recours à différents juges qui ont effectués le travail individuellement sans consultation entre eux. Puisque les données provenaient d'une autre étude, le choix des juges avait été fait préalablement. Pour les banques de questions 1 et 2, la qualité de chacune des questions n'a été évaluée que par un juge alors que pour la banque de questions 3, deux juges ont évalué chacune des questions. Les juges se sont distribué la tâche de façon égale entre les différentes banques. Ainsi, les consignes portant sur le contenu ont été évaluées par deux résidents en médecine (experts de contenu) et les consignes portant sur la forme ont été évaluées par deux étudiants gradués en mesure et évaluation agissant comme experts de forme.

Analyses

Dans le but de vérifier si les consignes retenues étaient adaptées à l'intention d'estimer la qualité (respect ou non des consignes) des questions, nous avons réalisé une analyse d'items, selon les principes de la théorie classique des tests (Crocker et Algina [24]) pour vérifier la discrimination de chacune des consignes répertoriées. Les items mentionnés ici concernent en l'occurrence bien les consignes et non pas les questions d'examens. Les analyses sont résumées sur 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 provenant de chacune des banques de données. Autrement dit, l'appréciation du respect de chaque consigne a été appliquée aux 407 questions qui provenaient des trois banques de questions. Pour chacune des questions qui

respectait la consigne analysée, un score de 1 était attribué à la consigne (0 étant un non-respect). L'accord inter-juges, représentant le degré de concordance de l'évaluation au sein de la paire de juges, a été calculé pour chaque consigne pour vérifier la standardisation dans l'utilisation de la liste. Ce calcul a été fait uniquement pour la banque de questions 3 puisque deux juges ont évalué la qualité de chacune des questions. Les logiciels Excel (version 14) et SPSS 17.0 ont été utilisés pour les analyses.

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 en fonction du respect des consignes. 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 :

- deux coefficients de discrimination supérieurs à 0,10 (pour une consigne dans les deux banques) ou ;
- un coefficient de discrimination supérieur à 0,10 dans une des deux banques et l'accord inter-juges supérieur à 0,70 dans la banque 3 (classification selon Ebel et Frisbie [25] et critère de Fleiss [26]).

Cette dernière stratégie a été choisie car cela représente une source d'informations supplémentaire comme il existe certaines contradictions dans les deux banques de données relatives aux QCM. Nous estimons donc que, si la consigne est comprise par les deux juges qui ont évalué la qualité des questions de la banque 3, elle le sera probablement par les rédacteurs de questions ; les futurs utilisateurs se rapprochent des juges qui ont évalué la qualité des 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 II présente les moyennes des indices de discrimination des consignes pour chaque banque de questions, les moyennes d'accord inter-juges de chacune des consignes de rédaction ; il catégorise en outre les consignes selon qu'elles font référence à un type donné de questions (tous les types de questions, QCM, QROC) ou au contenu.

Consignes relatives aux QROC

Six consignes de rédaction associées aux QROC satisfaisaient le critère d'inclusion de coefficient de discrimination supérieur à 0,10 de la banque de questions 1. Les coefficients de discrimination des consignes, lorsqu'elles sont appliquées à des questions de type 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 par rapport à la qualité des questions lorsqu'elles sont appliquées à 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). Ainsi, lorsqu'elles sont respectées pour la rédaction des questions de type QROC, ces six consignes permettent plus facilement de distinguer les questions de moins bonne qualité de celles de meilleure qualité (discrimination).

Consignes relatives aux QCM

Pour les deux banques de données (banques 2 et 3), 16 consignes de rédaction de questions avaient, que ce soit dans la banque 2, 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 II.

Quatre consignes de rédaction ont été retenues, car leur coefficient de discrimination était supérieur à 0,10 dans les deux banques de questions. Il s'agit des consignes 5 (discrimination dans la banque 2 = 0,181 ; discrimination dans la banque 3 = 0,243), 13 (discrimination dans la banque 2 = 0,195 ; discrimination dans la banque 3 = 0,229), 17 (discrimination dans la banque 2 = 0,330 ; discrimination dans la banque 3 = 0,208) et 20 (discrimination dans la banque 2 = 0,155 ; discrimination dans la banque 3 = 0,229). De plus, l'accord inter-juges pour ces consignes variait de 0,95 à 1,00 dans la banque de questions 3. 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) sont créditées d'un coefficient de discrimination adéquat selon les critères de sélection (> 0,10), mais dans une seule des deux banques de données. Les coefficients de ces consignes étaient respectivement de 0,164 ; 0,175 ; 0,155 et 0,175 dans la banque de questions 3. Cependant, 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 par rapport à la qualité des questions. Ainsi, ces quatre consignes de rédaction (nommées ci-haut) ont pu être conservées en tenant compte de l'accord inter-juges (> 0,70). L'accord inter-juges se situait entre 0,95 et 1,00 pour chacune des quatre consignes.

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 fait que ces consignes répondaient aux critères de sélection, soit un coefficient de

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

Consignes de rédaction	Banque de questions 1	Banque de questions 2	Banque de questions 3	
	Indice de discrimination	Indice de discrimination	Indice de discrimination	Accord inter-juges
Applicables à tous les types de questions				
La question d'introduction doit être une phrase complète	0,042	0,074	0,042	1,00
L'énoncé de la question doit porter sur un seul problème, un seul contenu	0,042	−0,032	–	1,00
La question d'introduction doit être brève	−0,015	0,117‡	−0,045	0,95
Utiliser un vocabulaire approprié pour le niveau du répondant	*0,263*†	*−0,062*	–	*1,00*
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*†
Les directives doivent indiquer clairement s'il faut identifier la bonne ou la meilleure réponse	–	−0,273	0,318‡	0,41
Éviter les mots inutiles dans l'énoncé de la question	–	−0,003	0,134‡	0,98
Ne pas utiliser de formulation négative	*0,161*†	*0,026*	*−0,190*	*0,95*
Applicables aux QCM				
Utiliser les questions à choix multiples 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
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*†
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
Éviter les questions de Type K (choix de réponse complexes, ex. : a et b mais pas c)	N/A	−0,003	–	1,00
« 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*†
La longueur des choix de réponse doit être constante	N/A	0,067	−0,070	0,79
Les choix de réponse ne doivent pas être plus longs que l'énoncé de la question	N/A	−0,254	−0,089	0,93
Les choix de réponse ne doivent pas contenir de mots répétitifs lorsque possible	N/A	−0,093	0,293‡	1,00
Les choix de réponse doivent être mutuellement indépendants/exclusifs	*N/A*	*0,330*‡	*0,208*‡	*0,95*†
Le contenu des choix de réponse doit demeurer homogène	N/A	–	–	0,95
Éviter l'utilisation d'adverbes tels que parfois, quelques fois, toujours, jamais, etc	N/A	–	–	1,00
Éviter de pister la bonne réponse par des constructions grammaticales erronées	*N/A*	*0,155*‡	*0,229*‡	*0,98*†
Applicables aux QROC				
Rédiger la question de telle sorte qu'il n'existe qu'une seule bonne réponse	*0,195*†	*N/A*	*N/A*	*N/A*
Éviter les réponses qui dépassent une courte phrase	*0,287*†	*N/A*	*N/A*	*N/A*
Éviter les réponses qui comprennent plus de 6 éléments (mot, expression, etc.)	0,022	N/A	N/A	N/A
Indiquer le degré de précision attendu, lorsque pertinent	*0,145*†	*N/A*	*N/A*	*N/A*
Indiquer si l'insertion d'éléments non pertinents sera pénalisée	*0,138*†	*N/A*	*N/A*	*N/A*

Tableau II. (suite).

Consignes de rédaction	Banque de questions 1	Banque de questions 2	Banque de questions 3	
Applicables aux questions de contenu				
Utiliser des distracteurs plausibles ; éviter les distracteurs illogiques	−0,070	–	0,055	1,00
Créer chacune des questions à partir d'un objectif éducationnel	−0,094	–	–	1,00
Les choix de réponse doivent inclure seulement une réponse correcte	–	–	*0,175*‡	*1,00*†
Les exemples et les énoncés ne doivent pas provenir du *textbook*	–	–	–	1,00
Les distracteurs devraient inclure des erreurs communément commises par les étudiants	–	–	–	1,00
Éviter de rédiger des choix de réponse d'une façon trop technique	–	–	−0,115	0,31
Utiliser des expressions familières qui sont incorrectes dans les distracteurs	–	–	–	0,98
Utiliser des énoncés vrais qui répondent incorrectement à la question	–	–	0,155‡	0,67*
Éviter des questions ambiguës qui pourraient empêcher de bien répondre à la question	–	0,117‡	−0,020	0,67
Éviter de développer les questions nécessitant des connaissances trop spécifiques	–	–	*0,155*‡	*0,98*†
Éviter les questions créées à partir d'opinion	–	–	*0,175*‡	*1,00*†

Les consignes en italiques sont celles retenues dans les deux grilles ; N/A : Consigne non-applicable pour ce type de question.
† : 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.

Tableau III. Grille de consignes rédaction pour la qualité des questions à choix multiples (QCM).

Consignes	
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

discrimination >0,10 et un accord inter-juges >0,70, car leur coefficient de discrimination était négatif dans l'une ou l'autre des banques de questions (2 ou 3), indice d'une mauvaise compréhension de la consigne de la part des juges.

Au total, huit consignes de rédaction pour les QCM ont été identifiées comme étant discriminantes de la qualité des questions selon les critères de sélection (tableau III). Autrement dit, les questions de meilleure qualité étaient celles qui respectaient le plus de ces consignes. 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 (consigne 5), quatre portent sur la forme d'une QCM (consignes 10, 13, 17 et 20) et trois font référence au contenu de la question rédigée (consignes 28, 35 et 36).

Discussion

L'élaboration de questions de qualité pour les examens écrits n'est pas une tâche aisée pour les rédacteurs. Une des stratégies recommandables pour faciliter cette tâche est l'utilisation de consignes de rédaction, à la condition que

l'impact positif du respect de cette consigne ait été documenté. Grâce à cette étude, nous avons identifié 14 consignes de rédaction qui peuvent discriminer les questions de bonne qualité des moins bonnes pour des questions de types QCM et QROC. La qualité des questions sur laquelle la discrimination s'est basée a été déterminée en fonction du nombre de consignes respectées par la question. Ces consignes sont toutes opérationnelles, mutuellement exclusives et applicables au contexte de l'évaluation des apprentissages dans le champ 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, une majorité de consignes retenues pour les QCM font référence au contenu ou 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. En effet, il existe peu de données probantes à ce sujet. Dès lors, dans une perspective docimologique, les rédacteurs de questions peuvent se baser sur ces résultats, c'est-à-dire suivre ces consignes identifiées comme discriminant les questions de meilleure qualité, c'est-à-dire aux consignes dont le respect, au moment de la rédaction des questions, est associé à des indicateurs de qualités psychométriques considérés comme élevés. Sept consignes relatives aux QCM n'ont pas été retenues malgré le respect des critères de sélection que nous avions établis à la base (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 de ces banques de questions 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 (tableaux III et IV). 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 juges n'avaient reçu aucune formation quant à l'utilisation de ces consignes, ce qui représente la situation dans laquelle les rédacteurs de questions se retrouveront lorsqu'ils devront rédiger de nouvelles questions d'examens écrits.

Une autre retombée intéressante de l'utilisation de ces 14 consignes de rédaction est qu'elles représentent une alternative économique en terme de ressources (temps, coût et experts de contenu) comparativement 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 [27–28]. Elle combine les connaissances et habiletés des experts de contenu avec le pouvoir algorithmique d'un programme informatique pour créer de nouvelles QCM [29]. 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 ne le prétendent les auteurs.

Une limite de cette étude est que les questions d'examens écrits sélectionnées dans les trois banques étaient à la base de bonnes questions. Ces questions respectaient probablement déjà un certain nombre de consignes, en raison des processus rigoureux des projets de recherche ainsi que du programme préclinique de médecine de notre institution (Université de Sherbrooke). Ainsi, les questions utilisées dans le cadre des 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.

Tableau IV. Grille de consignes de rédaction pour la qualité des questions à réponse ouverte courte (QROC).

	Consignes
1	Utiliser un vocabulaire approprié pour le niveau du répondant
2	Éviter une 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

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 en fonction du respect des consignes de rédaction. Quatorze consignes de réaction (six pour les QROC et huit pour les QCM) ont ainsi été identifiées, donnant maintenant aux rédacteurs de questions 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.

Contributions

Élise Vachon Lachiver a mené cette étude dans le cadre d'une maîtrise en sciences cliniques avec un cheminement en pédagogie des sciences de la santé. Elle a participé à l'analyse des données, à l'interprétation des résultats, ainsi qu'à la rédaction et à la révision du manuscrit. Christina St-Onge a dirigé cette étude en contribuant à la conception du protocole de recherche ainsi qu'à l'analyse des données, à l'interprétation des résultats et à la révision critique, et à l'approbation de la version finale du manuscrit. Jacinthe Cloutier a participé à la collecte des données, la révision critique du manuscrit et à l'approbation de la version finale. Paul Farand a contribué à la conception de l'étude, à la révision critique du manuscrit et à l'approbation de sa version finale.

Approbation éthique

Compte tenu de la nature du travail de recherche, aucune autorisation d'un comité d'éthique n'a été sollicitée, conformément à la politique d'éthique des trois conseils de recherche du Canada.

Déclaration d'intérêts

Les auteurs déclarent n'avoir aucun lien d'intérêt concernant les données publiées dans cet article.

Remerciements et financements

Les auteurs remercient les organisations ayant octroyé des bourses contributives au financement de ce travail :

Fonds de développement pédagogique de la Société des médecins de l'Université de Sherbrooke (SMUS), concours 2011. Chercheur principal : Pr Paul Farand.

Fonds d'innovation pédagogique de l'Université de Sherbrooke, concours 2013–2014. Chercheure principale : Christina St-Onge.

Chaire de recherche en pédagogie médicale Paul Grand'Maison de la Société des Médecins de l'Université de Sherbrooke, bourse accordée dans le cadre de la maîtrise de Élise Vachon Lachiver.

Annexe: Références des consignes de rédaction répertoriées

Livre

Auger R, Séguin SP, Nézet-Séguin C. Formation de base en évaluation des apprentissages : Module 3, construction de l'instrument de mesure. Outremont (QC) : Les Éditions Logiques, 2000, 68 pages.

Osterlind SJ. Constructing test items : Multiple-choice, constructed-response, performance, and other formats (2e ed.). Boston (MA) : Kluwer Academic Publishers, 1998.

Recommandations de bonnes pratiques – guidelines révisé par les pairs

Frary RB. More multiple-choice item writing do's and don'ts. Pr Assess Res Eval 1995;4;11. [On-line] Disponible sur : http://pareonline.net/getvn.asp?v=4&n=11.

Hogan, TP, Murphy G. Recommendations for Preparing and Scoring Constructed-Response Items: What the Experts Say. Appl Meas Educ 2007;20;4:427-441.

Synthèse de connaissance révisée par les pairs

Frey BB, Petersen S, Edwards LM, Teramoto Pedrotti J, Peyton V. Item-writing rules: Collective wisdom. Teach Teach Educ 2005;21:357-364.

Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Appl Meas Educ 1989;2;1:37-50.

Revue systématique

Haladyna TM, Downing SM, Rodriguez MC. A Review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002;15;3:309-334.

Article de recherche original

Hansen JD, Dexter L. Quality multiple-choice test questions: Item-writing guidelines and an analysis of auditing testbanks. J Educ Bus 1997;73;2:94-97.

Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42: 198-206.

Références

1. Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners, 2002.
2. Brady AM. Assessment of learning with multiple-choice questions. Nurse Educ Pract 2005;5:238-42.
3. Epstein RM, Hundert EM. Professional Competence. JAMA 2002;287:226-35.
4. Linn RL, Gronlund NE. Measurement and Assessment in Teaching. Upper Saddle River, New Jersey: Merrill Prentice-Hall, 2000.
5. van der Vleuten C. Validity of final examinations in undergraduate medical training. BMJ 2000;321:1217-9.
6. Jouquan J. L'évaluation des apprentissages des étudiants en formation médicale initiale. Pédagogie Médicale 2002;3: 38-52.
7. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ 2003;37:830-7.

8. Rodriguez MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educ Meas Issues Pract 2005;24:3-13.
9. Bush ME. Quality assurance of multiple-choice tests. Qual Assur Educ 2006;14:398-404.
10. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ 2002;15:309-33.
11. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ 2009;9:40.
12. Caldwell DJ, Pate AN. Effects of question formats on student and item performance. Am J Pharm Educ 2013;77:1-5.
13. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. Adv Health Sci Educ Theory Pract 2005;10: 133-43.
14. Considine J, Botti M, Thomas S. Design, format, validity and reliability of multiple choice questions for use in nursing research and education. Collegian 2005;12:19-24.
15. Prihoda TJ, Pinckard RN, McMahan CA, Jones AC. Correcting for guessing increases validity in multiple-choice examinations in an oral and maxillofacial pathology course. J Dent Educ 2006;70:378-86.
16. Haladyna TM, Downing SM. A taxonomy of multiple-choice. Appl Meas Educ 1989;2:37-50.
17. Bertrand R, 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 RJ, Muñiz J. New guidelines for developing multiple-choice\nItems. Methodol Eur J Res Methods Behav Soc Sci 2006;2:65-72.
19. DiBattista D, Kurzawa L. Examination of the quality of multiple-choice items on classroom tests. Can J Scholarsh Teach Learn 2011;2:1-23.
20. Tarrant M, Ware J. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 2008;42:198-206.
21. DeVellis RF. Scale development: Theory and applications. Thousand Oaks: Sage publications, 2016.
22. 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.
23. McConnell MM, St-Onge C, Young ME. The benefits of testing for learning on later performance. Adv Health Sci Educ Theory Pract 2015;20:305-20.
24. Crocker L, Algina J. Introduction to classical and modern test theory. Sea Harbor Drive: Holt, Rinehart and Winston, 1986.
25. Frisbie DA, Ebel RL. Essentials of educational measurement. Upper Saddle River: Prentice Hall, 1991.
26. Fleiss JL. Balanced incomplete block designs for inter-rater reliability studies. Applied Psychological Measurement, Thousand Oaks: Sage publications, 1981.
27. Luecht RM. Adaptive Computer-based tasks under an assessment engineering paradigm. In : D. J. Weiss (Ed.). *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing. 2009* [On-line] Disponible sur : www.psych.umn.edu/psylabs/CATCentral/.
28. Gierl MJ, Zhou J, Alves C. Developing a taxonomy of item model types to promote assessment engineering. J Technol Learn Assess 2008 7:1-51.
29. Gierl MJ, Lai H. Evaluating the quality of medical multiple-choice items created with automated processes. Med Educ 2013;47:726-33.

Citation de l'article : Vachon Lachiver É., St-Onge C., Cloutier J., Farand P., La rédaction de questions à choix multiple et de questions à réponse ouverte et courte pour les examens écrits dans les programmes de formation en santé : une étude docimologique des consignes de rédaction. Pédagogie Médicale 2017:18;55-64