Modalité docimologiques de l'évaluation des compétences en second cycle – Options et optimisation

JP. Fournier

(Nice)

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

Secondes Rencontres Pédagogiques d'Ile de France
Paris 2020

Quel(s) Objectif(s) ?

I. Introduction

La réforme du 2e cycle repose sur l'approche par compétences et permettra la mise en place d'un curriculum basé sur une approche par compétences, comme c'est déjà le cas dans la quasi-totalité des autres filières de la santé. Une telle approche vise avant tout à aider les étudiants à agir efficacement en milieu professionnel.

Le nouveau programme du 2e cycle des études médicales repose sur trois piliers :

1. Les connaissances théoriques contextualisées,
2. L'apprentissage systématique du raisonnement clinique et à la résolution de problèmes en période de stage surtout (dont l'amorce est proposée par différentes « situations de départ » ou de référence)
3. Une valorisation du parcours de l'étudiant.

- Le paradigme d'apprentissage

L'approche par compétences implique de passer du paradigme d'enseignement au paradigme d'apprentissage. Ce qui est important, c'est ce que l'étudiant apprend et la façon dont il l'apprend. L'enseignant a alors pour rôles de guider, de soutenir et vérifier les apprentissages.

- Le développement du raisonnement clinique

Pour leur permettre d'agir efficacement, les connaissances des médecins doivent être organisées en réseau dans la mémoire à long terme, afin d'être mobilisées au bon moment et au bon endroit, dans le cadre d'un raisonnement clinique approprié. Les connaissances théoriques de ce programme doivent être mobilisables et donc leur mémorisation fixée par des contextes cliniques et par un apprentissage à la résolution de problèmes. Les enseignants guideront les étudiants, pas à pas en s'appuyant sur les connaissances apprises. C'est dire combien le lien entre « les situations de départ » et les items de connaissances est important dans la formulation des objectifs de connaissances.

- La professionnalisation des étudiants en médecine

La démarche de professionnalisation vise donc à transformer l'étudiant en professionnel de santé, en l'aidant à développer ses compétences, à construire son identité de médecin et à partager des valeurs et des normes communes à notre profession.

N° 3. Le raisonnement et la décision en médecine. La médecine fondée sur les preuves (Evidence Based Medicine, EBM). La décision médicale partagée. La controverse

- Analyser les principes du raisonnement hypothético déductif et de la décision contextualisée en médecine.
- Décrire la démarche EBM ; en préciser les limites.
- Apprécier dans chaque situation clinique, le poids respectif des trois types de données constituant une approche EBM.
- Préciser la notion de niveau de preuve dans son raisonnement et dans sa décision.
- Définir les notions d'incertitude et de controverse
- Identifier les circonstances d'une décision médicale partagée avec le patient et son entourage (voir item 322).
- Préciser les notions d'efficacité, d'efficience et d'utilité dans le raisonnement et la décision médicale.
- Comprendre et apprendre la notion de discussion collégiale pour les prises de décision en situation de complexité et de limite des savoirs.

Quels outils ?

List 1

Practical Guidance for Clinical Reasoning Assessment From a 2016 Scoping Review of Clinical Reasoning Assessment Methods

- Multiple assessment methods (i.e., non-WBAs, assessments in simulated clinical environments, and WBAs) should be used as part of a clinical reasoning assessment program.
- Many individual assessment methods can obtain adequate reliability for high-stakes assessment (≥ 0.8) with an adequate number of items or cases, broad sampling, and sufficient testing time.
- To ensure competence, a large number of assessments are needed, administered longitudinally, that cover a variety of clinical problems in diverse settings to accommodate content and context specificity.
- Methods should be chosen based on coverage of the different components of clinical reasoning, validity, feasibility, defensibility, and fit for the purpose of the assessment.
- Whole- and part-task assessment methods (i.e., those that cover all versus a few components of clinical reasoning) used together can ensure measurement of the whole construct *and* adequate sampling.
- Non-WBAs (e.g., MCQs, EMQs, KFEs) have the advantage of broad sampling, blueprinting, control, and consistency. They can also assess accuracy.
- MCQs and KFEs have the best validity evidence regarding content, internal structure, and consequences or outcomes on clinical practice performance; however, they have significant issues with cueing when it comes to response process.
- Non-WBAs measure a more limited number of components of clinical reasoning compared with simulations and WBAs, which tend to measure more of the whole task.
- WBAs are embedded in actual clinical practice, lending authenticity to content and response process validity; however, content coverage is not systematic.
- The defensibility of using WBAs for summative decisions is questionable because, from a generalizability theory perspective, a large number of measurements are needed to reach acceptable reliability for judgments. Ensuring evaluation by multiple raters over time is also essential for WBAs.
- Whole-task clinical reasoning assessments (i.e., those that cover the full range of tasks from information gathering to differential diagnosis to management and treatment) are essential for formative feedback and assessment for learning.
- Assessments in simulated clinical environments and WBAs are essential parts of any comprehensive assessment strategy because they ensure that learners are assessed on the whole task, though they are time- and resource-intensive to develop and administer.

- ❖ Nécessité de juxtaposer des évaluations hors contexte professionnel et en contexte professionnel réel ou reproduit (simulation) ;
- ❖ Pour les évaluations hors contexte professionnel :
- ✓ Utiliser ce qui existe déjà plutôt qu'inventer de nouveaux formats ;
- ✓ Nécessité d'utiliser plusieurs formats complémentaires pour balayer les différents facettes du raisonnement ;
- ✓ Formats valides (contenu, construit, prédictive) ;
- ✓ Niveau de fidélité (reproductibilité des scores) adéquat (≥ 0,8) pour les épreuves d'enjeu important ;
- ✓ Épreuves réalisables et défendables(recours) ;
- ✓ Impact didactique ;
- ✓ Approche pragmatique.

D'après

Daniel M. *Acad Med* 2019
Schuwirth LWT. *Med Educ* 2004
Lee M. *Med Teach* 2018
Sharma S. *J Grad Med Educ* 2019
Hrynchak P. *Med Educ* 2014
Tamblyn R. *JAMA* 2007
Tamblyn R. *Arch Intern Med* 2010
Huwendiek S. *Med Teach* 2017

Quels moyens ?

Assessment method: Definition	Clinical reasoning component						
	IG	HG	PR	DD	LD	DJ	MT
Non-workplace-based assessments							
Clinical or comprehensive integrative puzzles: An extended matching crossword puzzle designed to assess a learner's ability to relate clinical vignettes to specific diagnoses and diagnostic or therapeutic interventions.	0.4	0.3	0.6	1.1	1.9	0.4	1.3
Concept maps: A schematic method for learners to organize and represent their knowledge and knowledge structures through a graphical illustration of the complex processes and relationships between concepts within a subject domain.	0.4	0.4	1.2	1.0	0.4	0.8	0.9
Extended matching questions: A written exam format consisting of a lead-in question (clinical vignette) followed by multiple answer options in a list where more answer options are given than in multiple-choice questions (i.e., > 5).	0.2	0.3	0.2	0.8	1.7	0.3	1.3
Key feature examinations: Problems typically consist of a clinical vignette followed by 2–3 questions that assess the critical elements ("key features") or challenging decisions that clinicians must make.	0.9	0.5	0.4	1.5	1.4	0.6	1.4
Multiple-choice questions: A clinical vignette is followed by up to 5 alternatives. Questions may take the following formats: single best alternative, matching, true or false, and combinations of alternatives.	0.9	0.3	0.0	0.6	1.9	0.0	1.8
Modified essay questions: A method wherein serial information about a clinical case is presented chronologically. After each item, the learner must document a decision. The student cannot preview subsequent items until a decision is made.	1.3	1.2	1.0	1.6	1.7	1.3	1.7
Oral examinations: A verbal examination conducted by one or more faculty members through unscripted or semiscripted questions that assess clinical reasoning and decision-making abilities, as well as professional values.	1.3	1.3	1.1	1.8	1.8	1.9	1.9
Patient management problems: A clinical scenario is presented in real-life settings with specific resources available for diagnosis or management. The learner chooses among multiple alternatives. The results of actions (e.g., labs, images) are provided.	1.6	1.0	0.3	1.4	1.9	0.6	1.7
Script concordance tests: Clinical scenarios with uncertainty are followed by a series of questions (e.g., if you are thinking X and you find Y, the answer becomes more likely, less likely, or no change). Responses are compared with those of experts.	0.4	0.8	0.6	0.8	1.3	0.9	1.1
Short- or long-answer (essay) questions: A clinical vignette is followed by one or more questions. Learners provide free-text responses that range in length from a few words to several sentences.	0.8	1.2	1.2	1.8	1.7	1.8	1.7

IG : Information gathering
HG : Hypothesis generation
PR : Problem representation
DD : Differential diagnosis
LD : Leading diagnosis
DJ : Diagnosis justifications
MT : Management and treatment

D'après Daniel M. *Acad Med* 2019

Quels moyens ?

Format	Avantages (théoriques)	Inconvénients (réels)
Comprehensive Integrative Puzzle	Aborde les modalité du raisonnement	Très peu de données Non informatisable
Cartes conceptuelles	Aborde les modalité du raisonnement Remédiation	Evaluation formative seulement
Extended Matching Questions	Pallient les questions ouvertes Limitent le *cueing effect*	Construction délicate Utilisation marginale (USMLE) Nombre d'options non défini
Modified Essay Questions	Presque équivalent des DCP Mesurent des habiletés cognitive élevées	Non utilisables pour les examens à enjeux importants Discordance entre correcteurs Faible consistance interne et faisabilité limitée Pas mieux que les QCM bien construits
Patient Management Problems	Equivalent des dossiers type ECN 2004 Administrables par informatique Mesurent des habiletés cognitive élevées	Spécificité de contenu Rédaction délicate et longue Correction délicate Faible consistance interne et faisabilité limitée Scores corrélés avec les QCM Impact didactique variable
Short and Long Answer Questions	Equivalent des dossiers et « questions courtes » d'internat Mesurent des habiletés cognitive élevées Corrigeables par non-médecins (grille) Limitent le *cueing effect*	Correction délicate Très faible fidélité Inutilisable pour des examens à enjeu important Faible corrélation avec les QCM et les examens de patients

D'après Daniel M. *Acad Med* 2019

Quels moyens ?

Assessment method: Definition	Clinical reasoning component						
	IG	HG	PR	DD	LD	DJ	MT
Non-workplace-based assessments							
Clinical or comprehensive integrative puzzles: An extended matching crossword puzzle designed to assess a learner's ability to relate clinical vignettes to specific diagnoses and diagnostic or therapeutic interventions.	0.4	0.3	0.6	1.1	1.9	0.4	1.3
Concept maps: A schematic method for learners to organize and represent their knowledge and knowledge structures through a graphical illustration of the complex processes and relationships between concepts within a subject domain.	0.4	0.4	1.2	1.0	0.4	0.8	0.9
Extended matching questions: A written exam format consisting of a lead-in question (clinical vignette) followed by multiple answer options in a list where more answer options are given than in multiple-choice questions (i.e., > 5).	0.2	0.3	0.2	0.8	1.7	0.3	1.3
Key feature examinations: Problems typically consist of a clinical vignette followed by 2–3 questions that assess the critical elements ("key features") or challenging decisions that clinicians must make.	0.9	0.5	0.4	1.5	1.4	0.6	1.4
Multiple-choice questions: A clinical vignette is followed by up to 5 alternatives. Questions may take the following formats: single best alternative, matching, true or false, and combinations of alternatives.	0.9	0.3	0.0	0.6	1.9	0.0	1.8
Modified essay questions: A method wherein serial information about a clinical case is presented chronologically. After each item, the learner must document a decision. The student cannot preview subsequent items until a decision is made.	1.3	1.2	1.0	1.6	1.7	1.3	1.7
Oral examinations: A verbal examination conducted by one or more faculty members through unscripted or semiscripted questions that assess clinical reasoning and decision-making abilities, as well as professional values.	1.3	1.3	1.1	1.8	1.8	1.9	1.9
Patient management problems: A clinical scenario is presented in real-life settings with specific resources available for diagnosis or management. The learner chooses among multiple alternatives. The results of actions (e.g., labs, images) are provided.	1.6	1.0	0.3	1.4	1.9	0.6	1.7
Script concordance tests: Clinical scenarios with uncertainty are followed by a series of questions (e.g., if you are thinking X and you find Y, the answer becomes more likely, less likely, or no change). Responses are compared with those of experts.	0.4	0.8	0.6	0.8	1.3	0.9	1.1
Short- or long-answer (essay) questions: A clinical vignette is followed by one or more questions. Learners provide free-text responses that range in length from a few words to several sentences.	0.8	1.2	1.2	1.8	1.7	1.8	1.7

IG : Information gathering
HG : Hypothesis generation
PR : Problem representation
DD : Differential diagnosis
LD : Leading diagnosis
DJ : Diagnosis justifications
MT : Management and treatment

D'après Daniel M. *Acad Med* 2019

Utilisation des QCM

- Balayent largement les programmes (validité de contenu) ;
- Scores reproductibles (*reliability*) et discriminants ;
- Corrélation à d'autres formats d'évaluation ;
- Economiques à développer (???) et à administrer ;
- Impact éducatif positif : *progress tests*, *test-enhanced learning* (TEL) ;
- En partie prédictifs des *performances* ultérieures ;
- Mais :

✓ Mesurent surtout des connaissances (niveau 1 de la taxonomie de Blum) ;
✓ Impact des réponses au hasard ou du *cueing effect* ;
✓ Impact éducatif négatif.

D'après
Swanson D. *Acad Med* 1992
Blake JM. Acad Med 1996
Van der Vleuten CPM. *Med Educ* 2005
Dijkserhuis MGK Med Teacher 2009
Pugh D. *Med Teacher* 2019
Norman G. *Med Teacher* 2010
Green M. *Med Teacher* 2018
Lee M. Med teach 2018
Sharma S. *J Grad Med Educ* 2019
Holmboe ES. *Arch Intern Med* 2008
Wenghofer E. *Med Educ* 2009
Schuwirth LW. *Med Educ* 1996
Downing SM. *Adv Health Sci Educ* 2005
Newble DI. *Med Educ* 1983

QCM : fidélité et faisabilité

Table 1 Reliability estimates of different assessment instruments as a function of testing time

Instrument	Description	Reliability for different testing times			
		1 hour	2 hours	4 hours	8 hours
Multiple choice*[42]	Short stem and short menu of options	0.62	0.76	0.93	0.93
Patient management problem*[42]	Simulation of patient, full scenarios	0.36	0.53	0.69	0.82
Key feature case (write-in)*[43]	Short patient case vignette followed by write-in answer	0.32	0.49	0.66	0.79
Oral examination†[44]	Oral examination based on patient cases	0.50	0.69	0.82	0.90
Long case examination†[45]	Oral examination based on previously unobserved real patient	0.60	0.75	0.86	0.90
OSCE*[46]	Simulated realistic encounters in round robin format	0.54	0.69	0.82	0.90
Mini-clinical exercise (mini-CEX)‡[47]	Short follow-up oral examination based on previously observed real patient	0.73	0.84	0.92	0.96
Practice video assessment†[16]	Selected patient–doctor encounters from video recordings in actual practice	0.62	0.76	0.93	0.93
Incognito standardised patients‡[48]	Real consultations scored by undetected simulated patients 0.86	0.61	0.76	0.82	0.86

* One-facet all random design with items crossed with persons (pxi).
† Two-facet all random design with judges (examiners) nested within items within persons (j:i:p).
‡ One-facet all random design with items nested within persons (i:p).

UNIVERSITÉ CÔTE D'AZUR
FACULTÉ DE MÉDECINE

D'après Van der Vleuten CPM. *Med Educ* 2005

Les QCM c’est nul !

Quel(s) examen(s) devez-vous prescrire chez une femme de 28 ans consultant aux urgences pour des douleurs de la fosse iliaque droite ? (QRM)

A- dosage de béta-HCG

B- dosage de CRP

C- dosage de procalcitonine

D- dosage de TP-TCA-fibrinogène

E- réalisation d’un hémogramme

Quelle est la problématique ?

Intérêt d'une vignette (QCM à contexte riche) ?

Une telle approche est-elle validée par la plus-value amenée en terme de niveau d'expertise attendue, en terme de discrimination, en terme de fidélité ?

Au moins sur le versant fidélité

TABLE 1
Reliability and Validity of Internal Medicine and Subspecialty Examinations

Examination	*N*	*Item Type*	*Number of Items*	*KR20*	*Adjusted Reliability*[a]	*Criterion-Related Validity*[b]	*Corrected Criterion-Related Validity*[c]
Internal medicine							
1991	8669	MCQ	142	.87	.74	.44	.47
		MTF	254	.91	.91	.43	.45
1992	9053	MCQ	148	.88	.75	.48	.51
		MTF	257	.90	.89	.44	.46
1992 Subspecialties							
Clinical cardiac electrophysiology	560	MCQ	73	.77	.73	.28	.32
		MTF	266	.84	.82	.18	.19
Hematology	452	MCQ	115	.82	.71	.32	.36
		MTF	378	.89	.84	.33	.35
Infectious disease	570	MCQ	135	.85	.71	.42	.45
		MTF	347	.88	.83	.32	.34
Nephrology	587	MCQ	112	.81	.66	.31	.34
		MTF	296	.87	.80	.36	.39
Pulmonary disease	1068	MCQ	127	.84	.66	.34	.38
		MTF	383	.79	.79	.29	.31
Rheumatology	387	MCQ	159	.80	.68	.27	.29
		MTF	251	.85	.82	.26	.29

MCQ : QCM à contexte riche, type A ;
MTF : QCM vrai/faux type QRM

D'après Downing SM. *Appl Meas Educ* 1995

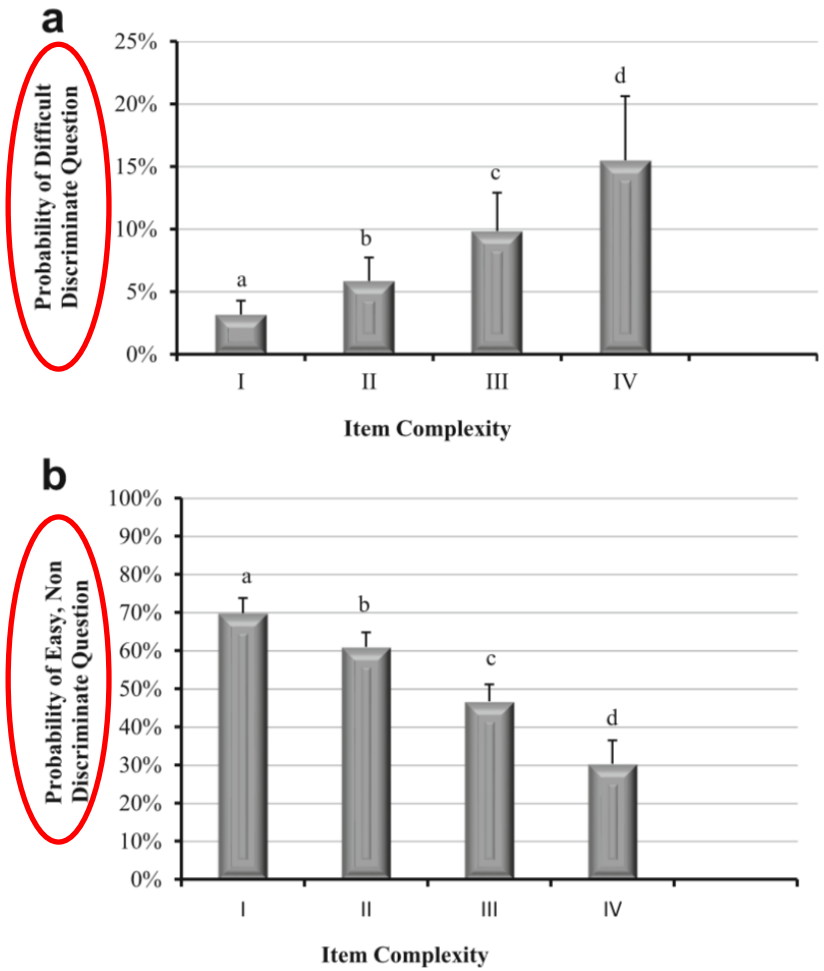
Peut-on mresurer plus que des connaissances déclaratives ?

Classification de Bloom*	Concepts	Sciences de la Santé
Mémoriser	Extraire des connaissances pertinentes de la mémoire à long terme	Question de mémorisation pure (cours)
Comprendre	Déterminer la signification d'informations orales, écrites ou de schémas	Physiopathologie Reconnaissance de structures sur imagerie
Appliquer	Exécuter ou utiliser une procédure dans une situation donnée	Prise de décision : diagnostic, thérapeutique, prise en charge, etc…
Analyser	Décomposer les parties constitutives d'un tout et déterminer les liens qui unissent ces parties entre elles et à une structure ou une finalité d'ensemble	Identification de données pertinentes Justification de décision
Evaluer (synthétiser initialement)	Porter un jugement sur la base de critères de normes	Synthèse de la vignette clinique (?)
Créer	Assembler des éléments pour former un tout nouveau et cohérent ou faire une production originale	NA

*Révisée par Anderson (2001)

D'après Scully D. *Pract Ass Res Eval* 2017

Pour quel intérêt ?



120 QCM de sémiologie administrés en DFASM 1, 2 et 3 en ligne, dont 34 Bloom 1 (0,28) :

Score global (/100) :

DFASM 1 : 60,78 ± 8,24

DFASM 2 : 66,15 ± 6,24 — $p < 0,001$

DFASM 3 : 76,25 ± 6,44 — $p < 0,001$

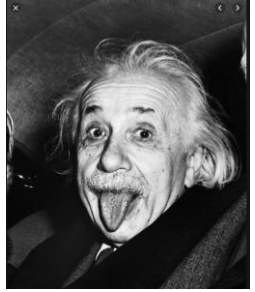
Pour les 34 QCM Bloom 1 (/100) :

DFASM 1 : 56,55 ± 9,46

DFASM 2 : 59,54 ± 6,87 — $p = 0,003$

DFASM3 : 61,33 ± 6,59 — $p = 0,104$

D'après Rush BR. *BMC Med Educ* 2016



Modèle d'approche cognitive

- Directement issu de l'intelligence artificielle (génération automatique de questions) ;
- Principes :

✓ Identification des éléments pertinents nécessaires à la résolution d'un problème ;

✓ Traduits en variables utilisées dans la vignette clinique et les options de réponses ;

✓ Guident les étudiants selon des directions pré-déterminées (réponse exacte ***et*** distracteurs) ;

- Limitation du *cueing-effect* ;
- Probablement beaucoup plus efficaces dans l'évaluation des processus cognitifs évolués que les QCM traditionnels.

D'après Lai H. *Teach Learn Med* 2016
Pugh D. Med Teach 2019

Modèle d'approche cognitive

Vous voyez aux urgences une femme de 28 ans que les pompiers amènent pour un malaise survenant alors qu'elle se levait pour aller aux toilettes. « Je me suis levée et la tête s'est mise à tourner. J'ai cru que j'allais m'évanouir » précise-t-elle. Elle se plaint depuis le matin de douleurs de la fosse iliaque droite d'intensité croissante, non calmées par le tramadol qu'elle a pris. « En plus, j'ai mes règles que je n'attendais pas » ajoute-t-elle. Elle a vomi 2 fois depuis le matin. Elle n'a pas d'antécédent en dehors d'une appendicite traitée médicalement il y a 2 mois. Elle est pâle. La PA est à 85-60 mm Hg, le pouls à 110 bpm, la température auriculaire à 37,8 ° C.

Quel est le premier examen à prescrire parmi les suivants ? (**QRU**)

A- dosage de béta-HCG

B- dosage de CRP

C- dosage de procalcitonine

D- dosage de TP-TCA-fibrinogène

E- réalisation d'un hémogramme

Modèle d'approche cognitive

Vous voyez aux urgences une femme de 28 ans que les pompiers amènent pour un malaise survenant alors qu'elle se levait pour aller aux toilettes. « Je me suis levée et la tête s'est mise à tourner. J'ai cru que j'allais m'évanouir » précise-t-elle. Elle se plaint depuis le matin de douleurs de la fosse iliaque droite d'intensité croissante, non calmées par le tramadol qu'elle a pris. « En plus, j'ai mes règles que je n'attendais pas » ajoute-t-elle. Elle a vomi 2 fois depuis le matin. Elle n'a pas d'antécédent en dehors d'une appendicite traitée médicalement il y a 2 mois. Elle est pâle. La PA est à 85-60 mm Hg, le pouls à 110 bpm, la température auriculaire à 37,8 ° C.

Quel est le premier examen à prescrire parmi les suivants ? (**QRU**)

A- dosage de béta-HCG

B- dosage de CRP

C- dosage de procalcitonine

D- dosage de TP-TCA-fibrinogène

E- réalisation d'un hémogramme

Vous voyez aux urgences une femme de 28 ans que les pompiers amènent pour un malaise survenant alors qu'elle se levait pour aller aux toilettes. « Je me suis levée et la tête s'est mise à tourner. J'ai cru que j'allais m'évanouir » précise-t-elle. Elle se plaint depuis le matin de douleurs de la fosse iliaque droite d'intensité croissante, non calmées par le tramadol qu'elle a pris. « En plus, j'ai mes règles que je n'attendais pas » ajoute-t-elle. Elle a vomi 2 fois depuis le matin. Elle n'a pas d'antécédent en dehors d'une appendicite traitée médicalement il y a 2 mois. Elle est pâle. La PA est à 85-60 mm Hg, le pouls à 110 bpm, la température auriculaire à 37,8 ° C.

Quel est le premier examen à prescrire parmi les suivants ? (**QRU**)

A- dosage de béta-HCG

B- dosage de CRP

C- dosage de procalcitonine

D- dosage de TP-TCA-fibrinogène

E- réalisation d'un hémogramme

Qu'est ce qu'un « bon » distracteur ?

- ❖ Vraisemblable ;
- ❖ Choisi par au moins 5 p. cent des étudiants ;
- ❖ Discriminant :
- ✓ Index de discrimination ≥ 0,25 à 0,3 ;
- ✓ Corrélation bisérielle de points (*point biserial*) positive ;
- ❖ Qu'indique SIDES ? :

Enoncé général

Quel examen allez-vous prescrire une fois traitée la patiente dans les heures qui viennent ?

Afficher les effectifs

Réponse	Proposition	Taux de coches	Sup	Inf	Disc	Discrimination
Fausse	écho-endoscopie œsophagienne	6.34%	5%	9%	5%	Insuffisant
Fausse	endoscopie œso-gastro-duodénale	45.07%	33%	53%	20%	Faible
Valide	TDM œsophagienne avec ingestion de produit de contraste hydrosoluble	26.76%	40%	9%	31%	Faible
Fausse	TDM œsophagienne injecté	21.83%	21%	28%	6%	Insuffisant

Fonctionne ?

Fonctionne ?

Rétro-action

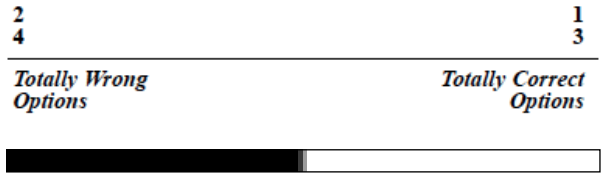
D'après Tavakol M. *Med Teacher* 2011

QRM ou QRU ?

QCM de type QRM :

- type vrai-faux ;
- 2-3 options correctes maximum :

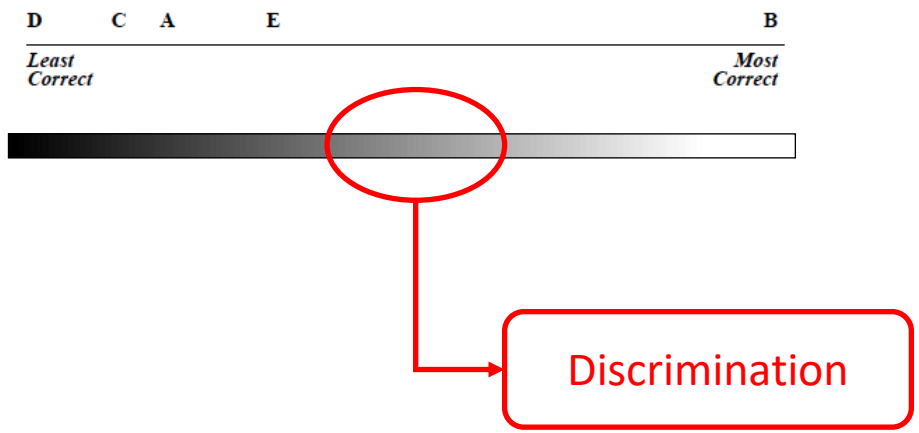
QCM vrai/faux :



QCM de type QRU (*one best answer*) :

- 1 seule réponse, non contestable ;
- 3 -4 options ***vraisemblables*** :

QCM *one best answer* :



D'après Case S. Swanson D. Constructing written questions for the basic and clincal sciences, NBME, Philadelphia, 2011

QCM ou QRU ?

5- Vous voyez en consultation un homme de 56 ans pour hémoptysie de moyenne abondance. Il est d'origine tchetchène, réfugié politique, sans domicile fixe. Un de ses coreligionnaires qui l'accompagne et sert d'interprète vous signale qu'il tousse et crache depuis plusieurs semaines, qu'il est très fatigué et a beaucoup maigri récemment. Il ajoute qu'ils sont plusieurs dans le squat à « cracher du sang et maigrir beaucoup ». Il précise que votre patient poursuit une intoxication alcoolo-tabagique ancienne. Vous disposez du cliché ci-dessous, réalisé la veille :

Quelle est votre prochaine prescription à visée diagnostique ? (**QRU**)

A- dosage de C Reactive Protein (CRP)
B- dosage d'interféron gamma (Quantiféron®)
C- réalisation d'un angioscanner thoracique
D- réalisation d'une bronchofibroscopie
E- réalisation de recherche de BAAR dans l'expectoration

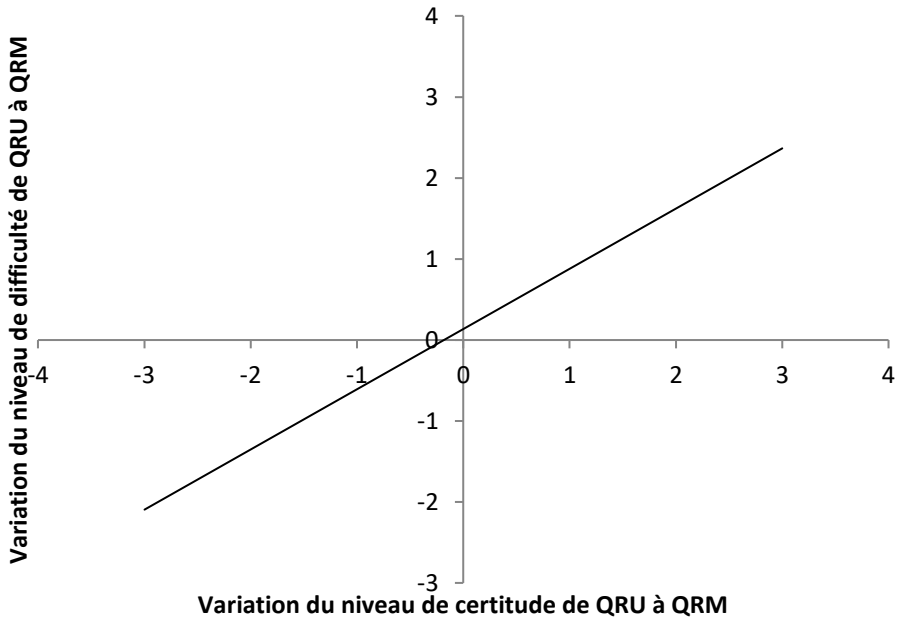
5- Vous voyez en consultation un homme de 56 ans pour hémoptysie de moyenne abondance. Il est d'origine tchetchène, réfugié politique, sans domicile fixe. Un de ses coreligionnaires qui l'accompagne et sert d'interprète vous signale qu'il tousse et crache depuis plusieurs semaines, qu'il est très fatigué et a beaucoup maigri récemment. Il ajoute qu'ils sont plusieurs dans le squat à « cracher du sang et maigrir beaucoup ». Il précise que votre patient poursuit une intoxication alcoolo-tabagique ancienne. Vous disposez du cliché ci-dessous, réalisé la veille :

Quelle(s) est(sont) votre(os) prochaine(s) prescription(s) à visée diagnostique ? (**QRM**)

A- dosage de C Reactive Protein (CRP)
B- dosage d'interféron gamma (Quantiféron®)
C- réalisation d'un angioscanner thoracique
D- réalisation d'une bronchofibroscopie
E- réalisation de recherche de BAAR dans l'expectoration

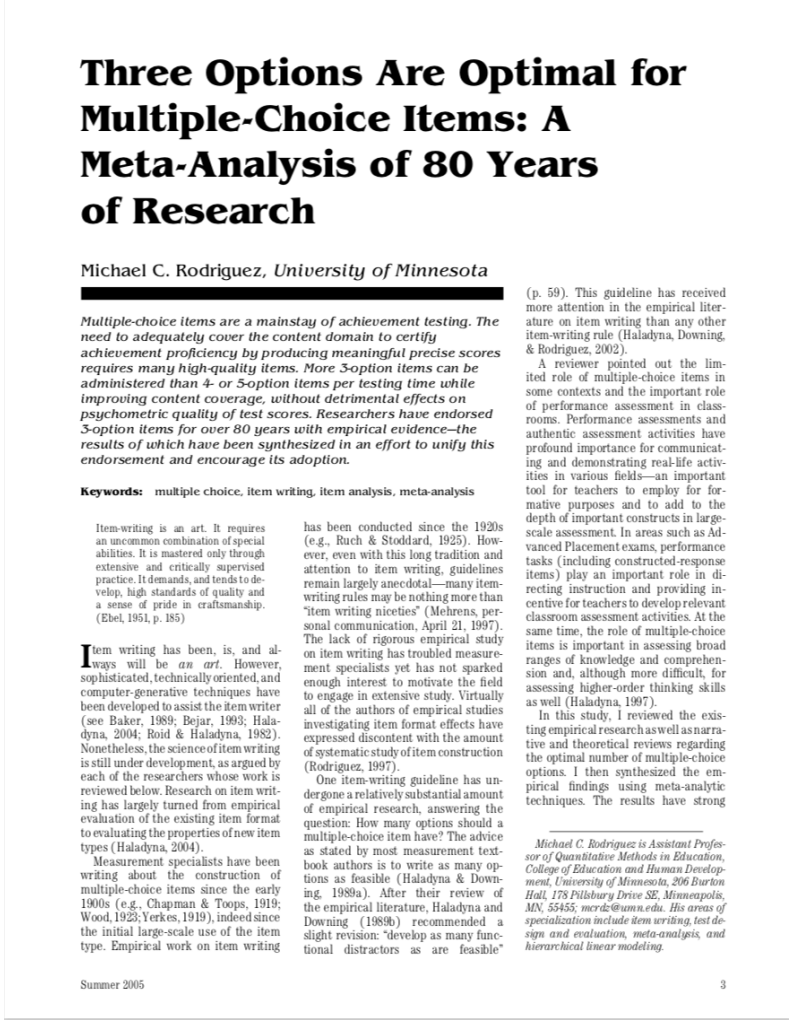
QCM ou QRU ?

Paramètres	Question	Résultats	p
Difficulté	QRU QRM	2,95 ± 0,92 2,76 ± 0,88	0,005
Certitude	QRU QRM	3,08 ± 1,05 2,79 ± 1,00	0,0003
Stratégie adaptée	QRU QRM	0,53 0,45	0,04



$r_s = 0,2923$,
$p < 0,0027$

Combien d'options de réponse ?



Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years of Research

Michael C. Rodriguez, *University of Minnesota*

Multiple-choice items are a mainstay of achievement testing. The need to adequately cover the content domain to certify achievement proficiency by producing meaningful precise scores requires many high-quality items. More 3-option items can be administered than 4- or 5-option items per testing time while improving content coverage, without detrimental effects on psychometric quality of test scores. Researchers have endorsed 3-option items for over 80 years with empirical evidence—the results of which have been synthesized in an effort to unify this endorsement and encourage its adoption.

Keywords: multiple choice, item writing, item analysis, meta-analysis

> Item-writing is an art. It requires an uncommon combination of special abilities. It is mastered only through extensive and critically supervised practice. It demands, and tends to develop, high standards of quality and a sense of pride in craftsmanship. (Ebel, 1951, p. 185)

Item writing has been, is, and always will be *an art*. However, sophisticated, technically oriented, and computer-generative techniques have been developed to assist the item writer (see Baker, 1989; Bejar, 1993; Haladyna, 2004; Roid & Haladyna, 1982). Nonetheless, the science of item writing is still under development, as argued by each of the researchers whose work is reviewed below. Research on item writing has largely turned from empirical evaluation of the existing item format to evaluating the properties of new item types (Haladyna, 2004).

Measurement specialists have been writing about the construction of multiple-choice items since the early 1900s (e.g., Chapman & Toops, 1919; Wood, 1923; Yerkes, 1919), indeed since the initial large-scale use of the item type. Empirical work on item writing has been conducted since the 1920s (e.g., Ruch & Stoddard, 1925). However, even with this long tradition and attention to item writing, guidelines remain largely anecdotal—many item-writing rules may be nothing more than "item writing niceties" (Mehrens, personal communication, April 21, 1997). The lack of rigorous empirical study on item writing has troubled measurement specialists yet has not sparked enough interest to motivate the field to engage in extensive study. Virtually all of the authors of empirical studies investigating item format effects have expressed discontent with the amount of systematic study of item construction (Rodriguez, 1997).

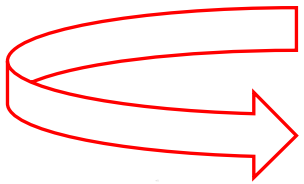
One item-writing guideline has undergone a relatively substantial amount of empirical research, answering the question: How many options should a multiple-choice item have? The advice as stated by most measurement textbook authors is to write as many options as feasible (Haladyna & Downing, 1989a). After their review of the empirical literature, Haladyna and Downing (1989b) recommended a slight revision: "develop as many functional distractors as are feasible" (p. 59). This guideline has received more attention in the empirical literature on item writing than any other item-writing rule (Haladyna, Downing, & Rodriguez, 2002).

A reviewer pointed out the limited role of multiple-choice items in some contexts and the important role of performance assessment in classrooms. Performance assessments and authentic assessment activities have profound importance for communicating and demonstrating real-life activities in various fields—an important tool for teachers to employ for formative purposes and to add to the depth of important constructs in large-scale assessment. In areas such as Advanced Placement exams, performance tasks (including constructed-response items) play an important role in directing instruction and providing incentive for teachers to develop relevant classroom assessment activities. At the same time, the role of multiple-choice items is important in assessing broad ranges of knowledge and comprehension and, although more difficult, for assessing higher-order thinking skills as well (Haladyna, 1997).

In this study, I reviewed the existing empirical research as well as narrative and theoretical reviews regarding the optimal number of multiple-choice options. I then synthesized the empirical findings using meta-analytic techniques. The results have strong

Michael C. Rodriguez is Assistant Professor of Quantitative Methods in Education, College of Education and Human Development, University of Minnesota, 206 Burton Hall, 178 Pillsbury Drive SE, Minneapolis, MN, 55455; mcrdz@umn.edu. His areas of specialization include item writing, test design and evaluation, meta-analysis, and hierarchical linear modeling.

…Oui, mais pas pour l'importance des réponses par « devinette » (*guessing*)



3 à 5 options de réponse

D'après Rodriguez MC. *Educ Meas* 2005

Impact du degré de certitude aux QCM

2 épreuves de 30 QCM (*one best answer*) en neurologie :

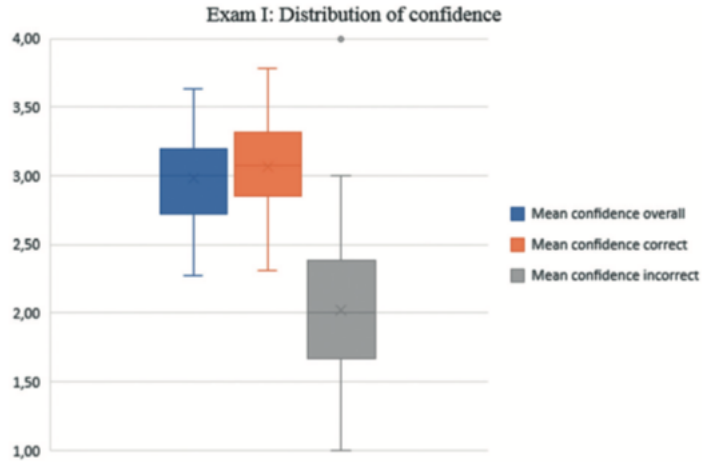
- 90 et 81 étudiants ;
- Mesure de l'impact clinique potentiel : bénin, risqué, dangereux ;
- Degré de confiance des étudiants : de très peu certain (1) à très certain (4) ;
- 4 niveaux de combinaisons :

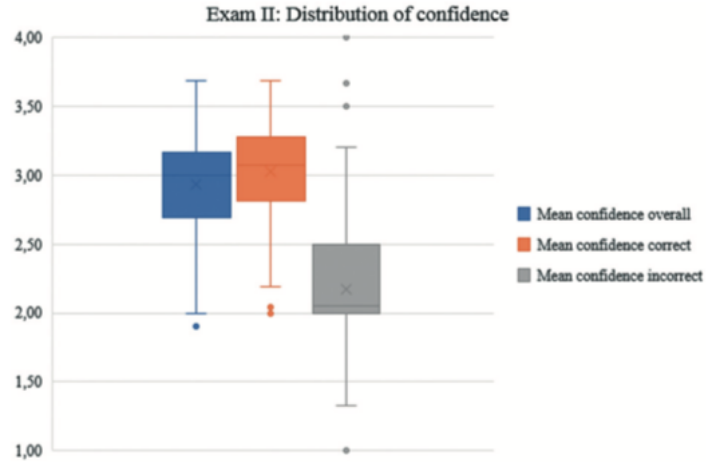
✓ Etudiants performants :	certain :	« informé » ;
	incertain :	« devinent » ;
✓ Etudiants non performants :	certain :	« mal informé » ;
	incertain :	« ignorant » ;

D'après Rangel RH. *Med Teacher* 2017

Impact du degré de certitude aux QCM

	Etude I	Etude II
Effectifs :	90	81
Taux de réussite :	0,91	0,87
Cronbach alpha :	0,71	0,61
Performance par QCM : Informés (%) : Devinent (%) : Mal informés (%) : Ignorants (%) :	 67,3 23,5 2,6 6,5	 70,3 26,1 3,9 8,1





D'après Rangel RH. *Med Teacher* 2017

Impact du degré de certitude aux QCM

	Etude I		Etude II	
Réponses correctes	Bénin :	3,14 [1,50 – 4,00]	Bénin :	3,12 [2,00 – 3,88]
	Risqué :	3,08 [2,29 – 3,85]	Risqué :	3,03 [2,00 – 3,80]
	Dangereux :	3,05 [2,00 – 3,92]	Dangereux :	2,96 [1,89 – 3,88]*
Réponses incorrectes	Bénin :	1,08 [1,00 – 3,00]	Bénin :	1,78 [1,00 – 3,00]
	Risqué :	1,42 [1,00 – 2,44]	Risqué :	1,98 [1,00 – 3,50]*
	Dangereux :	2,01 [1,50 – 2,57]*	Dangereux :	2,57 [1,00 – 4,00]*

Importance de la rétro-action

D'après Rangel RH. *Med Teacher* 2017

Finalement

- ❖ Le stimulus (vignette clinique) est l'élément déterminant, même plus important que les options de réponse ;
- ❖ La qualité des distracteurs fait la qualité du QCM ;
- ❖ Réserver les QRM aux questions de type vrai-faux ;
- ❖ Et au point de vue cognitif ?
- ✓ Utilisation des 2 modes de raisonnement complémentaires ;
- ✓ Les experts génèrent plus d'hypothèses que les étudiants ;
- ✓ En raisonnement à voix haute, les étudiants raisonnent différemment devant des QCM de type vrai-faux (niveau 1) et des QCM à contexte riche (niveaux 2 à 4) ;
- ✓ En IRM fonctionnelle, les différents modes de raisonnement correspondent à des circuits fonctionnels différents.

D'après

Coderre SP. *BMC Med Educ* 2004
Schuwirth LW. *Med Educ* 2004
Tarrant M. *Med Teach* 2008
Rush BR. *BMC Med Educ* 2016
Heist BS. *J Grad Med Educ* 2014
Surry LT. *Med Educ* 2017
Durning SJ. *Med Teach* 2016

Intérêt des DCP ?

- Aucun équivalent dans la littérature internationale ;
- Proche des *Modified-Essay-Questions* ;
- Très peu de données dans la littérature francophone ;
- Techniquement : possibilité d'utiliser des formes séquentielles de QCM, TCS et KFP ;
- Données locales :

✓ Faiblement mais significativement corrélés à la performance des internes de phase socle :
r = 0,264, p = 0,0234 ;

✓ Déterminants de la corrélation :

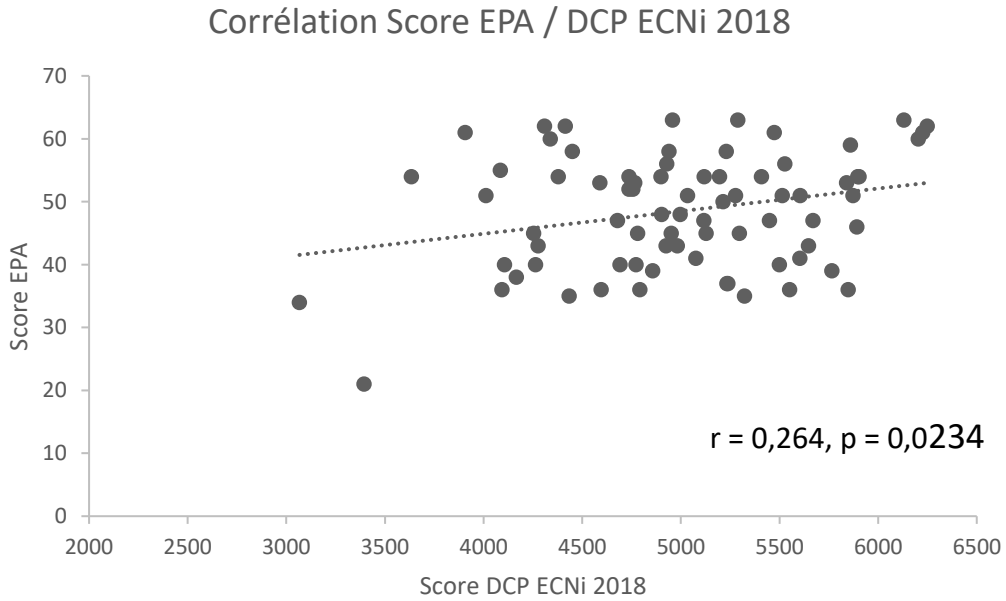
➢ Score de qualité : 2,72 ± 0,81 vs 2,29 ± 0,53 : p < 0,001 ;

➢ Proportion de questions de mémorisation : 0,28 *vs* 0,38, p = 0,15 ;

➢ Nombre moyen de question du programme par DCP : 6 ± 3 *vs* 3 ± 2, p = 0,07 ;

➢ Nombre moyen de spécialités différentes par DCP : 4 ± 2 *vs* 2 ± 1 : p = 0,148 ;

➢ Nombre moyen de questions : 14 ± 1 *vs* 15 ± 1, p = 0,08



D'après Case S. Swanson D. Constructing written questions for the basic and clincal sciences, NBME, Philadelphia, 2011
Huwendiek S. *Med Teach* 2017
Cooke S. *Med Teach* 2017

Qu'attendre des TCS ?

Vous vous apprêtez à partir en voiture le 11 février 2013. Il neige. Votre voiture est garée dans la rue. Elle ne démarre pas.

Si vous pensiez…	…et qu'alors vous constatez	L'impact de cette nouvelle information sur votre hypothèse est…
Que votre réservoir est vide	Vous avez roulé la veille avec le témoin de réserve du réservoir allumé	-2 -1 0 +1 +2
Que votre batterie est morte	Vous avez acheté votre voiture, neuve, il y a 4 ans	-2 -1 0 +1 +2
Que votre alternateur ne fonctionne pas	Rien ne s'allume au démarrage	-2 -1 0 +1 +2
-2 : Totalement négatif -1 : Négatif	0 : ni plus, ni moins positif	+1 : Positif +2 : Très positif

- Bases cognitives très solides : scripts, intuition, analyse ;
- Format docimologiquement très robuste ;
- Permet d'explorer des aspects peu ou pas mesurables :
- rapport bénéfice-risque, rapport coût-efficacité, éthique, incertitude ;
- Mesure le cheminement décisionnel ;
- Validité de construit ;
- Impact didactique (approche par concordance) ;
- Corrélation à l'évaluation in-situ ;
- Mais :

D'après :

- ✓ (Très) dépendant de la rigueur de sa construction ;
- ✓ Mode cognitif des experts et des étudiants mal connu ;
- ✓ Doutes sur validité et reproductibilité des scores ;
- ✓ Composition du jury

Lubarsky S. *Med Educ* 2011
Charlin B. *Med Educ* 2006
Charlin B. *Teach Learn* Med 2002
Lubarsky S. *Med Teach* 2013
Cooke S. *Acad Med* 2017
Helou MA. *Acad Med* 2020
Kelly W. *Teach Learn Med* 2012
Lubarsky S. *Perspect Med Educ* 2018
Lineberry M. *Med Educ* 2013
Gagnon R. *Adv in Health Sci Educ Theory Pract* 2011
Ramaekers S. *Ass Eval Higher Educ* 2010
Williams RG. *Acad Med* 2011
Gagnon R. *Adv Health Sci Educ* 2009
Gawad N. *Acad Med* 2020
Humbert A. *Med Teach* 2011
Duggan P. *BMC Med Educ* 2012

A 35-year-old mother of 3 presents to your office at 17.00 hours with complaints of severe, watery diarrhoea. On questioning, she indicates that she has been ill for about 24 hours. She has had 15 watery bowel movements in the past 24 hours, has been nauseated, but not vomited. She works during the day as a cook in a longterm care facility but left work to come to your office. On her chart, your office nurse notes a resting blood pressure of 105/50 mmHg supine (a pulse of 110/minute), 90/40 standing, and an oral temperature of 36·8 °. On physical examination, you find she has dry mucous membranes and active bowel sounds. A urinalysis (urine microscopy) was normal, with a specific gravity of 1·030.

1 What clinical problems would you focus on in your immediate management of this patient? List up to 3

2 How should you treat this patient at this time? Select up to 3

1 Antidiarrhoeal medication
2 Antiemetic medication
3 Intravenous 0·9% NaCl
4 Intravenous 2/3–1/3
5 Intravenous gentamicin
6 Intravenous metronidazole
7 Intravenous Ringer lactate
8 Nasogastric tube and suction
9 Nothing by mouth
10 Oral ampicillin
11 Oral chloramphenicol
12 Oral fluids
13 Rectal tube
14 Send home with close follow-up
15 Surgical consultation
16 Transfer to hospital

3 After management of the patient's acute condition, what additional measures, if any, would you take? Select up to 4 or select #11, none, if none are indicated

1 Avoid dairy products
2 Colonoscopy
3 Enteric precautions
4 Gastroenterology consultation
5 Give immune serum globulin to patients at longterm care facility
6 Infectious disease consultation
7 Notify Public Health Authority
8 Stool cultures
9 Strict isolation of patient
10 Temporary absence from work
11 None

Qu'attendre des KFP ?

- Focalisent sur les points importants et critiques ;
- Bases psychométriques très solides ;
- Solide expérience internationale ;
- Impact éducatif important ;
- Validité de construit ;
- Validité prédictive.

D'après

Page G. *Acad Med* 1995
Hrynchak P. *Med Teach* 2014
Farmer EA.*Med Educ* 2005
Fisher MR. *Med Teach* 2005
Bronander KA. *Med Teach* 2005
Huwendiek S *Med Te*ach 2017

En conclusion

- **Aucun format ne peut « tout faire »** ;
- 3 formats docimologiquement corrects ;
- 3 formats ***complémentaires*** :

✓ Même approche des étudiants les plus performants que les cliniciens entraînés aux QCM à contexte riche ;

✓ Mêmes erreurs des étudiants les moins performants sur des QCM à contexte riche que dans la « vraie vie » ;

✓ Les QCM représentent l'aboutissement du raisonnement (prise de décision) ;

✓ Les options des QCM, des KFP et des TCS correspondent à la phase d'intuition (génération des hypothèses pertinentes) ;

✓ Les TCS mesurent le cheminement du raisonnement (phase analytique) ;

✓ Les formes les plus évoluées de QCM à contexte riche permettent de représenter le problème (Bloom 5) et de mesurer le cheminement clinique (identification d'informations pertinentes – Bloom 4) ;

✓ Les KFP permettent une approche « en profondeur » avec un impact didactique ;

✓ KFP et QCM à contexte riche sont prédicteurs de la performance des futurs internes.

D'après Van der Vleuten CPM. *BMJ* 2000
Evans JSBT *Pers Psychol Sci* 2013
Rangel RH. *Med Teach* 2017
Heist BJ. *J Grad Med Educ* 2014
Surry LT. *Med Educ* 2017
Lubarsky S. *Med Teach* 2013
Krathwol DR. Theory into Pract 2002
Daniel M. *Acad Med* 2019
Hrynchak P. *Med Educ* 2014
Huwendiek S. *Med Teach* 201
Lee M. *Med Teach* 2018
Sharma S. *J Grad Med Educ* 2019

Merci de votre attention