

GUIDE PRATIQUE

POUR L'ÉLABORATION D'UNE ÉVALUATION PAR QUESTIONS À CHOIX MULTIPLE (QCM)

Ophélie DE CLERCQ, avril 2021



TABLE DES MATIÈRES



QCM : DÉFINITION(S) ET VARIANTES	3
Qu'est-ce qu'une QCM ?	3
De quoi est composée une QCM ?	3
Les différentes formes de QCM	4
La question à choix unique (QCU)	4
La question à réponses multiples (QRM)	5
Le Vrai/Faux	5
La QCM avec solutions générales (SG)	6
La QCM avec degré de certitude (DC)	6
La question à appariement (QAA)	7
La question d'ordonnement (QO)	7
Le texte à trous	8
POURQUOI ET QUAND CHOISIR LES QCM ?	9
Quels sont ses intérêts et ses limites pédagogiques ?	9
Les qualités attendues d'une évaluation	11
Choisir les QCM, dans quel but ?	11
Choisir les QCM, en lien avec quels acquis et dispositif d'apprentissage ?	12
Choisir les QCM, pour quel(s) niveau(x) de performance ?	13
QUELS TYPES DE QUESTIONS POUR DÉPASSER LA RESTITUTION ?	15
QCM et hauts niveaux d'acquisition	15
Utiliser des questions basées sur des situations concrètes	16
Utiliser des distracteurs de haute qualité	16
Utiliser des éléments visuels à analyser	18
Utiliser des questions qui favorisent la pensée « multi-logique »	18
Utiliser des questions de type « assertion-raison »	18
COMMENT RÉDIGER UNE QCM	20
La formulation des consignes	20
La formulation de l'énoncé	21
La formulation des solutions proposées	22
BARÈME(S) DE CORRECTION ET CORRECTION DU HASARD	24
Le barème classique (sans pénalisation)	24
Le barème utilisant la correction pour devinette	25
Le barème relevant le score minimal à atteindre	26
Le barème utilisant les degrés de certitude	28
Autres pistes pour diminuer les effets du hasard	29
COMMENT ANALYSER ET INTERPRÉTER LES RÉSULTATS ?	30
La prise en compte de la difficulté des questions	31
La prise en compte du pouvoir discriminant des items	31
L'analyse de la qualité des distracteurs utilisés	31
Attention à la constante macabre !	32
COMMENT FOURNIR UN FEEDBACK APPROPRIÉ AUX ÉTUDIANTS ?	33
RESSOURCES BIBLIOGRAPHIQUES	34





QCM : DÉFINITION(S) ET VARIANTES

I Qu'est-ce qu'une QCM ?

Selon Leclercq (1986), une **question à choix multiple (QCM)** est « une question à laquelle l'étudiant répond en opérant une sélection (au moins) parmi plusieurs solutions proposées, chacune étant jugée (par le constructeur de l'épreuve et par un consensus entre spécialistes) correcte ou incorrecte indépendamment de l'étudiant qui doit y répondre » (p. 15).

I De quoi est composée une QCM ?

Une question à choix multiple est composée de trois parties (Leclercq, 1986) :

- › **La consigne** écrite qui précise le type de questions, le mode de réponse, les principes de notation et le barème ; la consigne est généralement donnée en début d'épreuve pour l'ensemble des questions.
- › **L'amorce** qui décrit le problème et pose la question : phrase affirmative, interrogative, graphe, image, situation-problème, etc.
- › **Les solutions proposées** comprenant la ou les **solution(s) correcte(s)** et les **distracteurs** (ou leurres), c'est-à-dire les solutions incontestablement fausses.

I Les différentes formes de QCM

Parmi les formes d'évaluations écrites, on distingue classiquement les **questions ouvertes** (à réponses construites) et les **questions fermées** (à réponses choisies). Comme le proposent Bouvy & Warnier (2011), l'expression « question à choix multiple » recouvre en pratique différentes formes de questions à réponses fermées :

- › Question à réponse unique (QRU).
- › Question à réponses multiples (QRM).
- › Vrai/Faux (V/F).
- › QCM avec solutions générales (SG).
- › QCM avec degrés de certitude (DC).
- › Question à appariement (QAA).
- › Question d'ordonnancement (QO).
- › Textes à trous (avec propositions de réponses).

La question à choix unique (QCU)

La QCM classique comprend une amorce et des solutions dont une seule est correcte.

EXEMPLE DE QUESTION À CHOIX UNIQUE

(Les cahiers du Louvain Learning Lab n°10, 2020) :

QUELLE EST L'HÉPATITE LA PLUS FRÉQUEMMENT OBSERVÉE TROIS SEMAINES APRÈS UN VOYAGE EN AFRIQUE DU NORD ?

- ☐ a. L'hépatite B
- ☐ b. L'hépatite C
- ☐ c. L'hépatite A
- ☐ d. L'hépatite Delta

Une variante répandue consiste à demander à l'étudiant de sélectionner la proposition incorrecte. De manière générale, les formulations négatives sont à éviter (Haladyna & Downing, 1989a in Burton & al, 1991), car elles compliquent inutilement le traitement que l'étudiant doit opérer pour trouver la bonne réponse et font appel à des facteurs (attention, compétence linguistique) qui n'ont généralement rien à voir avec les objectifs visés.

De plus, le fait que l'étudiant trouve la réponse incorrecte ne garantit pas qu'il connaisse la réponse correcte.

En fonction des objectifs poursuivis, un item négatif peut cependant s'avérer approprié dans certains cas particuliers (par ex : des objectifs en lien avec des questions de sécurité). Il est alors préférable d'inclure la négation dans l'énoncé (plutôt que dans les solutions) et de bien mettre la négation en évidence en la soulignant ou en utilisant des lettres capitales.

Une autre variante de ce type de QCM consiste à demander aux étudiants de choisir l'option la plus prioritaire ou la plus pertinente dans le contexte. Autrement dit, certaines propositions sont incorrectes, d'autres sont correctes, mais une seule est clairement plus correcte ou pertinente que les autres dans le contexte décrit. C'est ce qu'on appelle une question à choix unique avec « meilleure réponse » ou « *best-answer item* » en anglais. Comme nous le verrons plus loin, il peut s'agir d'une variante intéressante à explorer pour élaborer des distracteurs à haut niveau de discrimination (en prenant garde évidemment de ne pas tomber dans la subjectivité – il doit y avoir consensus d'experts sur la meilleure option).

La question à réponses multiples (QRM)

Dans une question à réponses multiples, l'étudiant doit identifier la ou les propositions correctes.

EXEMPLE DE QUESTION À RÉPONSES MULTIPLES :

UNE GAZOMÉTRIE ARTÉRIELLE EST EFFECTUÉE EN AIR AMBIANT ET RETROUVE : PH 7,47 ; PAO260 MMHG ; PACO232MMHG ; HCO3- 23MMHG. SUR LA BASE DE CES GAZ DU SANG VOUS CONCLUEZ QUE (PLUSIEURS RÉPONSES CORRECTES) :

- ☐ a. La patiente est en acidose respiratoire
- ☐ b. La patiente est en alcalose respiratoire*
- ☐ c. La patiente est en acidose métabolique
- ☐ d. La patiente est en alcalose métabolique
- ☐ e. La patiente a une hypoxémie*

Le nombre de réponses correctes peut être annoncé ou non à l'étudiant, en fonction des acquis d'apprentissage évalués et du niveau de complexité attendu.

Par contre, il est important de toujours préciser le barème utilisé, selon que l'on accorde des points sur base de la connaissance totale ou partielle des étudiants. Dans le premier cas de figure, l'étudiant doit avoir choisi toutes les bonnes réponses pour avoir le point ; s'il coche un ou plusieurs distracteurs, il obtient zéro. Dans le deuxième cas de figure, l'attribution des points se fait indépendamment pour chaque alternative. L'étudiant obtient une fraction des points pour chaque bonne réponse cochée et pour chaque distracteur non coché.

Selon Burton & al (1991), les deux méthodes de notation sont sources de difficulté. Dans le cas du tout-ou-rien, l'étudiant qui coche tous les bonnes réponses sauf une obtient au final le même nombre de points (zéro) que l'étudiant qui n'en a trouvé aucune. Dans le cas de la valorisation partielle, l'étudiant qui ne connaît rien mais sélectionne toutes les réponses peut potentiellement obtenir autant de points qu'un étudiant dont la connaissance est partielle ; dans ce cas précis, il est donc recommandé de pénaliser les réponses incorrectes.

Pour terminer, il faut garder à l'esprit que les QRM sont nettement plus difficiles à traiter pour les étudiants que les QCU, chaque proposition constituant en quelque sorte une « sous-question ». Ceci augmente le temps nécessaire par question, mais peut également avoir une incidence importante sur les résultats des étudiants (en particulier si l'on ne récompense pas la connaissance partielle).

Le Vrai/Faux

L'énoncé d'un Vrai/Faux ne comporte qu'une seule solution correcte parmi deux propositions (vrai/faux, oui/non, exact/inexact, etc.).

EXEMPLE DE QUESTION DE TYPE « VRAI-FAUX » :

DANS UN ATOME DE CHARGE ÉLECTRIQUE NEUTRE, LE NOMBRE D'ÉLECTRONS EST ÉGAL AU NUMÉRO ATOMIQUE :

- ☐ a. Vrai*
- ☐ b. Faux

Comme pour les autres types, les réponses doivent être indubitablement correctes ou incorrectes, et poser un défi suffisant pour que la réponse ne puisse pas se trouver trop facilement sur Internet. Notons qu'il s'agit d'une forme moins fiable que la QCM classique, puisque l'étudiant a 50% de chance de trouver la bonne réponse en la sélectionnant au hasard. En outre, l'étudiant peut tout à fait savoir qu'une proposition est fausse sans nécessairement connaître la bonne réponse.

La QCM avec solutions générales (SG)

Les solutions générales sont des solutions génériques qui s'appliquent à tout le questionnaire. Dans une QCM avec solutions générales, « l'étudiant est averti qu'en plus des solutions propres à chaque question (k), il doit prendre en considération x autres solutions » (Leclercq, 1986).

A titre d'exemple, voici un système à quatre solutions générales :

- k+1 : aucune des solutions proposées n'est correcte ;
- k+2 : toutes les solutions proposées sont correctes ;
- k+3 : il manque au moins une donnée dans l'énoncé pour pouvoir répondre ;
- k+4 : une absurdité dans l'énoncé rend la question sans objet.

EXEMPLE DE QUESTION AVEC DEUX SOLUTIONS GÉNÉRALES (Cours Biologie générale, L. Ladrière, 2018):

PARMI LES PROTISTES SUIVANTS, LESQUELS SONT FLAGELLÉS ?

- ☐ a. Les rhizopodes
- ☐ b. Les actinopodes
- ☐ c. Les mastigophores*
- ☐ d. Les acrasiales
- ☐ e. Toutes les propositions sont correctes
- ☐ f. Aucune des propositions n'est correcte

Les solutions générales peuvent être explicites si elles figurent parmi les propositions, comme dans l'exemple ci-dessus. Elles peuvent également être implicites : dans le cas de questions à 4 propositions de réponse a, b, c, d par exemple, l'étudiant se voit alors préciser dans les consignes qu'il peut toujours répondre e, f, g ou h.

Ces solutions générales sollicitent des processus mentaux qui vont au-delà de la simple reconnaissance. Elles diminuent l'effet du hasard puisqu'elles augmentent le nombre de solutions. Les questions deviennent évidemment dès lors un peu plus complexes.

A noter que la solution générale « aucune des propositions n'est correcte » ne permet pas d'évaluer si l'étudiant connaît la bonne réponse. Il s'agit du même inconvénient que dans les vrai-faux ou dans les QCM avec item négatif, vus précédemment.

La QCM avec degré de certitude (DC)

La QCM avec degré de certitude tend également à réduire l'effet du hasard et a comme avantage de « pousser l'étudiant à évaluer sa propre compétence et lui apprend à assumer les conséquences des décisions qu'il prend » (Swinen, 2008, in Lepage & Romainville, 2009). Concrètement, il est demandé à l'étudiant d'associer à sa réponse un degré de certitude, c'est-à-dire de déterminer, sur une échelle, son niveau de confiance dans la réponse choisie. Ce degré de certitude intervient ensuite dans le barème de notation (voir plus bas, « barème utilisant les degrés de certitude »).

EXEMPLE DE QCM AVEC DEGRÉ DE CERTITUDE (Morvan, Clerquin, Leclercq, 2014 in Uyttebrouck, 2020) :

QUE S'EST-IL PASSÉ AU COURS DE LA PÉRIODE DITE « ÂGE PRÉ-AGRICOLE » PAR RAPPORT AUX PROCÉDÉS DE CONSERVATION DES ALIMENTS ? VEUILLEZ CHOISIR UNE RÉPONSE :

- a. Maîtrise des principes de conservation des aliments
- b. Découverte des grands principes de la conservation des aliments
- c. Développement de machines pour maîtriser les procédés
- d. Domestication et diffusion des techniques
- e. Mécanisation des techniques de conservation

ESTIMEZ LE DEGRÉ DE CERTITUDE QUI CORRESPOND LE MIEUX À VOTRE RÉPONSE À LA QUESTION PRÉCÉDENTE, EN CHOISISANT PARMIS CES PROPOSITIONS ::

- ☐ 0% de certitude
- ☐ 20% de certitude
- ☐ 40% de certitude
- ☐ 60% de certitude
- ☐ 80% de certitude
- ☐ 100% de certitude

L'intérêt pédagogique des QCM avec degré de certitude est qu'ils invitent à l'étudiant à évaluer ses propres connaissances. Selon Leclercq & Poumay (2007), cette forme d'évaluation est particulièrement formative,

car elle fait « *vivre à l'apprenant la gestion de son doute, de sa connaissance partielle, éléments accompagnant, tout processus d'apprentissage* ». Pour l'enseignant, les degrés de certitude offrent une vue beaucoup plus fine sur ce qui est effectivement maîtrisé en permettant de distinguer les connaissances réellement utilisables (réponses correctes avec haut degré de certitude), les connaissances inutilisables (réponses correctes ou incorrectes avec faible degré de certitude), ou encore les connaissances « dangereuses » (réponses incorrectes avec haut degré de certitude).

En vertu de l'alignement pédagogique (voir plus bas), il est fortement recommandé que les étudiants aient l'occasion de s'exercer à ce type de questions avant toute utilisation certificative des degrés de certitude.

La question à appariement (QAA)

Il s'agit d'une question dans laquelle l'étudiant doit mettre en relation des éléments issus de deux listes distinctes : une liste de « questions » et une liste de « solutions ». Par exemple, l'énoncé : « Appariez la capitale à son pays » proposera deux listes : « Canada, Italie, Japon » et « Ottawa, Rome, Tokyo ». Il existe deux formes de questions à appariement : la **correspondance univoque et la forme asymétrique par excès de solutions**.

Dans la première forme, les deux listes comportent le même nombre d'éléments. Si l'on reprend notre exemple des pays à appairer avec les capitales, cela signifie qu'il suffit d'associer correctement les deux premiers éléments pour trouver automatiquement la dernière réponse.

En raison de cet inconvénient, Leclercq (1986) suggère d'utiliser la **forme asymétrique par excès de solutions**. Dans ce cas de figure, il y a plus d'éléments dans une liste que dans l'autre, ce qui réduit le choix par déduction et la place laissée au hasard.

EXEMPLE DE QUESTION À APPARIEMENT ASYMÉTRIQUE PAR EXCÈS DE SOLUTIONS (Bouvy & Warnier, 2011) :

POUR CHACUN DES PATIENTS SUIVANTS, SÉLECTIONNEZ LE MÉDICAMENT QUI ENTRERA LE PLUS PROBABLEMENT DANS SON TRAITEMENT.

PATIENTS :

1. Homme de 49 ans vous appelant à son chevet pour fièvre brutale, douleur Hemi thoracique gauche rythmée par la respiration, frisson solennel, expectorations rouillées : on perçoit sur l'hémothorax gauche des crépitements et de la pectoriloquie.
2. Home de 68 ans consultant pour fièvre, toux, hémoptysies, amaigrissement, auscultation normale en dépit de l'observation d'une lésion excavée de l'apex gauche, entourée de nodules et d'infiltras flous.
3. Jeune sportif de 28 ans présentant parfois une toux dyspnéique et sibilante au décours de son cross hebdomadaire.

PROPOSITIONS :

- ☐ a. Amoxicilline
- ☐ b. Quinolones
- ☐ c. Rifampicine
- ☐ d. Bêtamimétique d'action brève
- ☐ e. Anticholinergique d'action longue
- ☐ f. Anticoagulants
- ☐ g. Diurétiques
- ☐ i. Théophylline-retard
- ☐ j. Oxygène

La question d'ordonnancement (QO)

Il s'agit d'une question dans laquelle l'étudiant doit ordonner une série de propositions présentées de manière aléatoire en fonction d'un mode d'organisation indiqué dans l'énoncé.

EXEMPLE DE QUESTION D'ORDONNANCEMENT (Bouvy & Warnier, 2011) :

CLASSEZ DANS L'ORDRE CHRONOLOGIQUE D'APPARITION LES 8 STADES DE DÉVELOPPEMENT DE LA VIGNE PRÉSENTÉS CI-APRÈS :

- ☐ a. Nouaison
- ☐ b. Véraison
- ☐ c. Sortie des feuilles
- ☐ d. Débourrement
- ☐ e. Arrêt de croissance
- ☐ f. Grappes visibles
- ☐ g. Floraison
- ☐ h. Maturité

Le texte à trous

Le texte à trous (« cloze test » en anglais) est un texte qui comprend des lacunes pour lesquelles l'étudiant doit choisir la réponse correcte parmi plusieurs propositions (il est possible également de demander aux étudiants de produire la réponse, sans suggérer de propositions donc, mais il ne s'agit plus alors de questions à choix). Dans le cas d'un texte à trous à réponses intégrées, les options proposées peuvent être plus nombreuses que les lacunes afin d'éviter les choix par déduction comme dans les questions à appariement. Elles peuvent également être utilisées plusieurs fois, ce qui doit impérativement être précisé dans la consigne le cas échéant.

EXEMPLE DE TEXTE À TROUS

 http://www.englishdaily626.com/grammar_cloze.php?001 :

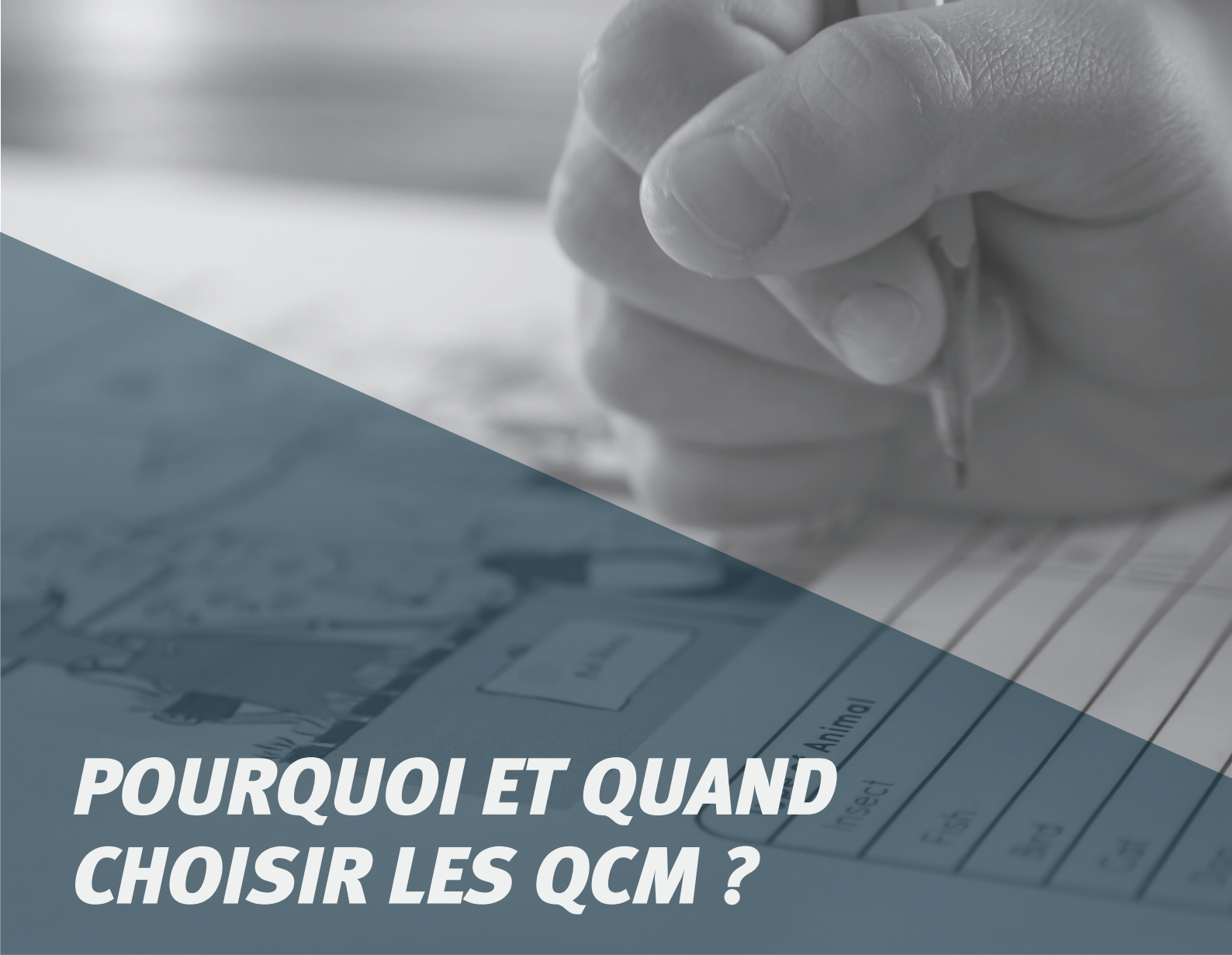
FILL IN EACH BLANK WITH THE CORRECT WORD FROM THE LIST. USE EACH WORD ONCE ONLY.

JOG – JOGS – JOGGED – TO JOG – ARE JOGGING – WILL JOG – JOGGING – IS JOGGING

Most days, Karen _____ alone but at times, when her father wants some exercise, he will join Karen on her runs. Karen loves _____ around her estate as there is beautiful greenery all around. At times she would _____ in the morning but because she usually works during the day, Karen goes _____ at night. Last week, Karen _____ all the way home from work! She really loves jogging.

Fréquent dans l'apprentissage des langues, le texte à trous peut également s'appliquer à d'autres domaines. En effet, les termes manquants peuvent être des mots de vocabulaire, des formes grammaticales, mais aussi des valeurs numériques, des dates, des concepts, des relations entre concepts (par ex. : inférieur/supérieur), etc.





POURQUOI ET QUAND CHOISIR LES QCM ?

Quels sont ses intérêts et ses limites pédagogiques ?

Les QCM sont largement utilisées dans l'enseignement supérieur pour mesurer les acquis d'apprentissage des étudiants ; comme toute méthode d'évaluation, les QCM présentent des forces et des faiblesses.

Leur intérêt principal est que **leur correction est rapide, simple et précise** et qu'elle peut être automatisée par lecture optique ou traitement informatique, ce qui en fait un outil privilégié pour **l'évaluation de grands groupes**. Les outils de traitement des résultats de QCM deviennent également de plus en plus performants, en offrant aux enseignants la possibilité d'avoir un feedback sur la qualité des questions (voir le chapitre sur l'analyse des items). Les QCM peuvent également être intégrées dans une **réflexion formative** en fournissant aux étudiants un retour sur leur degré de maî-

trise de la matière.

Un autre avantage des QCM est qu'il s'agit d'un outil d'évaluation relativement **objectif et fiable**. Toutes les réponses sont analysées de la même façon. Les QCM permettent en outre de tester **un plus large spectre** de matière que les autres formes d'évaluation pour lesquelles il est souvent impossible de poser des questions sur tous les chapitres d'un cours.

Malgré ces avantages évidents, construire de bonnes questions à choix multiple est moins simple qu'il n'y paraît. D'une part, il faut un nombre suffisant de questions pour balayer l'entièreté de la matière et, d'autre part, au sein de chaque question, il faut veiller à une **bonne formulation des questions**, mais aussi **des distracteurs** qui doivent être plausibles. Il est également préférable de tester ses questions auprès d'un collègue ou d'un assistant avant de les soumettre aux étudiants.



Le **temps de préparation** d'un examen par QCM est dès lors relativement conséquent, et la banque de questions doit être régulièrement renouvelée afin d'éviter la diffusion des questions entre étudiants d'année en année.

Un reproche qui est également adressé aux QCM est qu'elles **ne permettent pas** d'engager les étudiants dans des **processus cognitifs de haut niveau**. Autrement dit, elles permettent d'évaluer des capacités telles que la restitution, la compréhension, l'application ou l'analyse, mais pas les deux derniers niveaux de la taxonomie de Bloom (c.-à-d. « évaluer » et « créer »). Dans une QCM, l'étudiant est amené à sélectionner la solution correcte : on mesure ainsi sa capacité à reconnaître la bonne réponse, pas à la produire. Il ne permet donc pas d'évaluer des capacités de production (par ex. : rédaction, expression orale, etc.).

De plus, les QCM sont utiles pour évaluer des savoirs de type cognitif, mais **moins adaptés** à l'évaluation de **savoir-être** ou de **savoir-faire**. Pour ces diverses raisons, il est préférable que les QCM ne constituent pas le moyen exclusif pour vérifier les connaissances des étudiants au sein d'un programme de formation.

Pour terminer, une dernière limite pédagogique attribuée aux QCM est de laisser la **place au hasard** en offrant la possibilité à l'étudiant de sélectionner la bonne réponse sans la connaître. Comme le précise Leclercq (1986), il est parfois difficile de distinguer dans une QCM, les « choix heureux par ignorance » des « choix corrects par compétence ». Plusieurs procédés permettent néanmoins de compenser cet inconvénient tels que l'augmentation du nombre de propositions, le recours à des barèmes relevant le seuil de réussite, ainsi que l'utilisation de QCM avec degrés de certitude (nous y reviendrons dans le chapitre consacré à la notation).

TABLEAU 1. SYNTHÈSE DES AVANTAGES ET DES INCONVÉNIENTS DES QCM

+ LES AVANTAGES
Correction rapide qui peut être automatisée
Adapté à l'évaluation de grands groupes
Pour une évaluation formative ou sommative
Capacité à couvrir toute la matière
Modalité d'évaluation objective et fiable

— LES INCONVÉNIENTS
Temps de préparation conséquent
Difficulté de créer de bons distracteurs
Mobilisation de capacités cognitives de base
Peu adapté aux savoir-être et savoir-faire
Possibilité de répondre au hasard



Les qualités attendues d'une évaluation

Pour rappel, comme toute évaluation, une évaluation par QCM doit être :

- **pertinente** : il y a adéquation entre l'évaluation et les acquis d'apprentissage visés ;
- **valide** : il y a adéquation entre ce que je pense évaluer et ce que j'évalue réellement ;
- **fiable** : les résultats de l'évaluation seront identiques si elle est reproduite à un autre moment ou par un autre évaluateur ;
- **faisable** : la stratégie d'évaluation et sa mise en œuvre sont réalistes compte tenu des moyens ;
- **équitable** : la stratégie d'évaluation ne (dé)favorise aucun étudiant.

Choisir les QCM, dans quel but ?

Afin d'assurer la validité du processus d'évaluation, il convient de déterminer ce que l'on souhaite évaluer (les objectifs pédagogiques), de quelle façon (les méthodes d'évaluation), mais également dans quel(s) but(s). Les QCM peuvent être utilisés dans le cadre d'une évaluation diagnostique, formative ou certificative en fonction des objectifs que l'on souhaite atteindre.

- **Évaluation diagnostique** : l'évaluation diagnostique a pour fonction de dresser un bilan des connaissances avant une formation. Elle offre la possibilité à l'enseignant de prendre une décision anticipée pour orienter son action pédagogique et les phases d'apprentissage prévues.
- **Évaluation formative** : l'évaluation formative a pour fonction de favoriser la progression des apprentissages de l'étudiant en lui donnant un feedback sur ses erreurs pour qu'il puisse adapter ses stratégies d'apprentissage.

Cette forme d'évaluation est également utile à l'enseignant puisqu'elle lui fournit des indications sur les difficultés rencontrées.

Elle est effectuée en cours d'année, à une ou plusieurs reprises, et aucune note n'y est généralement associée. Pour des raisons d'alignement pédagogique (voir plus bas), l'évaluation formative doit évidemment préparer à l'évaluation certificative, et donc être de forme et de difficulté comparable. Les QCM se prêtent bien à de l'évaluation formative : si elles incluent des solutions commentées pour chaque item, elles peuvent être utilisées par les étudiants pour s'entraîner ou même s'auto-évaluer.

- **Évaluation certificative** : l'évaluation certificative ou sommative a pour fonction l'attestation ou la reconnaissance des apprentissages. Elle survient au terme d'un processus d'enseignement et elle sanctionne ou certifie le degré de maîtrise des apprentissages des étudiants par une note.

En résumé, on peut retenir pour la pratique que :

- Les QCM peuvent servir à orienter, réguler-améliorer ou valider les apprentissages et il importe de définir le but de l'évaluation au moment de la conception du dispositif pédagogique.
- Les QCM se prêtent bien à de l'évaluation formative, ainsi qu'à de l'auto-évaluation, à condition d'être correctement construites et d'inclure un feedback pour chaque item.
- Pour que l'évaluation formative prépare adéquatement les étudiants à l'évaluation certificative, les deux évaluations doivent être de forme et de difficulté similaires.

Choisir les QCM, en lien avec quels acquis et dispositif d'apprentissage ?

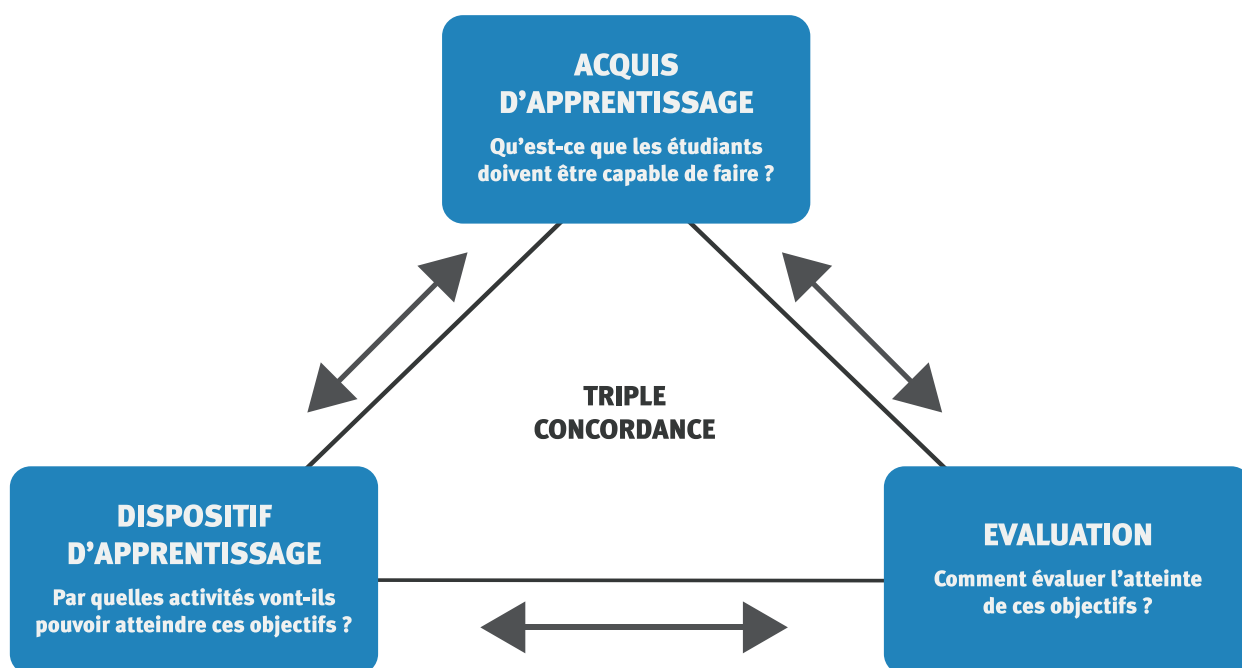
Il s'agit probablement de l'étape la plus cruciale dans la construction de l'évaluation. Rappelons que l'alignement pédagogique (également appelé « triple concordance ») est la congruence entre les acquis d'apprentissage (ce que l'étudiant doit savoir et être capable de faire à la fin du cours), le dispositif d'apprentissage (les activités proposées, les « méthodes pédagogiques ») et l'évaluation.

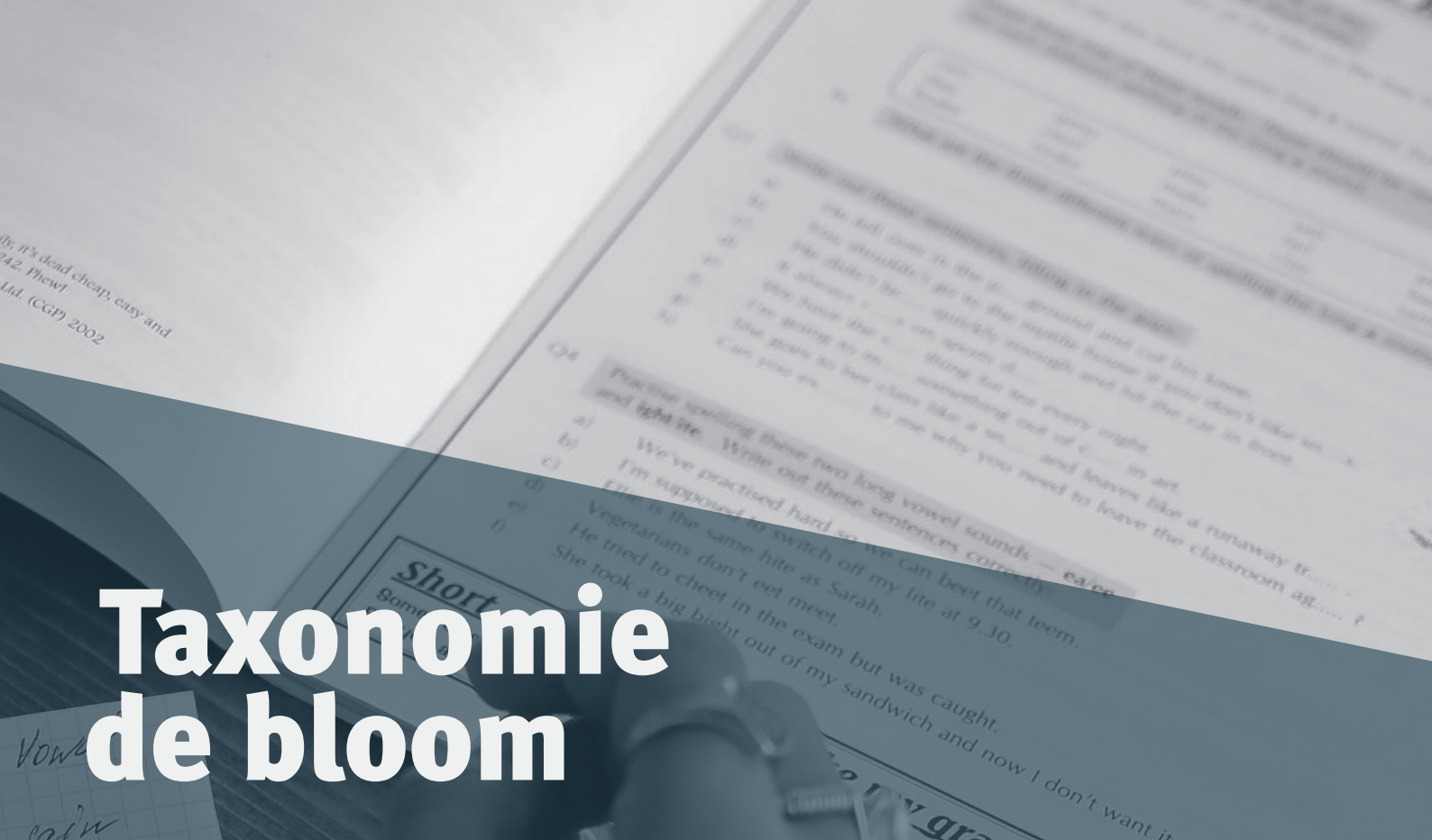
FIGURE 1. L'ALIGNEMENT PÉDAGOGIQUE (OU LA TRIPLE CONCORDANCE)

Le schéma ci-dessous synthétise bien la structure et la dynamique suggérée par la notion d'alignement pédagogique : le dispositif d'apprentissage doit être cohérent avec les acquis d'apprentissage visés, et l'évaluation cohérente avec les acquis comme avec le dispositif d'apprentissage.

En amont de la conception d'une épreuve par QCM (ou de tout autre type d'évaluation), il peut être utile de mettre à l'épreuve l'alignement pédagogique de l'enseignement concerné en passant en revue les questions suivantes (reprises de la fiche-outil du CAP sur la « triple concordance ») :

- Dans quelle mesure vos acquis d'apprentissage sont-ils explicites ? Pour vous ? Pour vos étudiants ?
- Y a-t-il concordance entre vos acquis d'apprentissage et le dispositif pédagogique que vous mettez en œuvre ?
Permettez-vous aux étudiants d'entraîner l'ensemble des compétences identifiées comme importantes dans vos intentions pédagogiques ?
Y a-t-il des compétences que vous n'entraînez pas ou d'autres sur lesquelles vous mettez l'accent, mais qui ne figurent pas dans vos acquis d'apprentissage ?
- Y a-t-il concordance entre vos acquis d'apprentissage et le type d'évaluation que vous mettez en œuvre ? Y a-t-il des compétences que vous n'évaluez pas ou d'autres sur lesquelles vous mettez l'accent, mais qui ne figurent pas dans vos acquis d'apprentissage ?
- Y a-t-il concordance entre votre dispositif pédagogique et l'évaluation que vous mettez en œuvre ? Y a-t-il des compétences que vous entraînez sans les évaluer ou, à l'inverse, des compétences que vous évaluez sans jamais les avoir entraînées ?





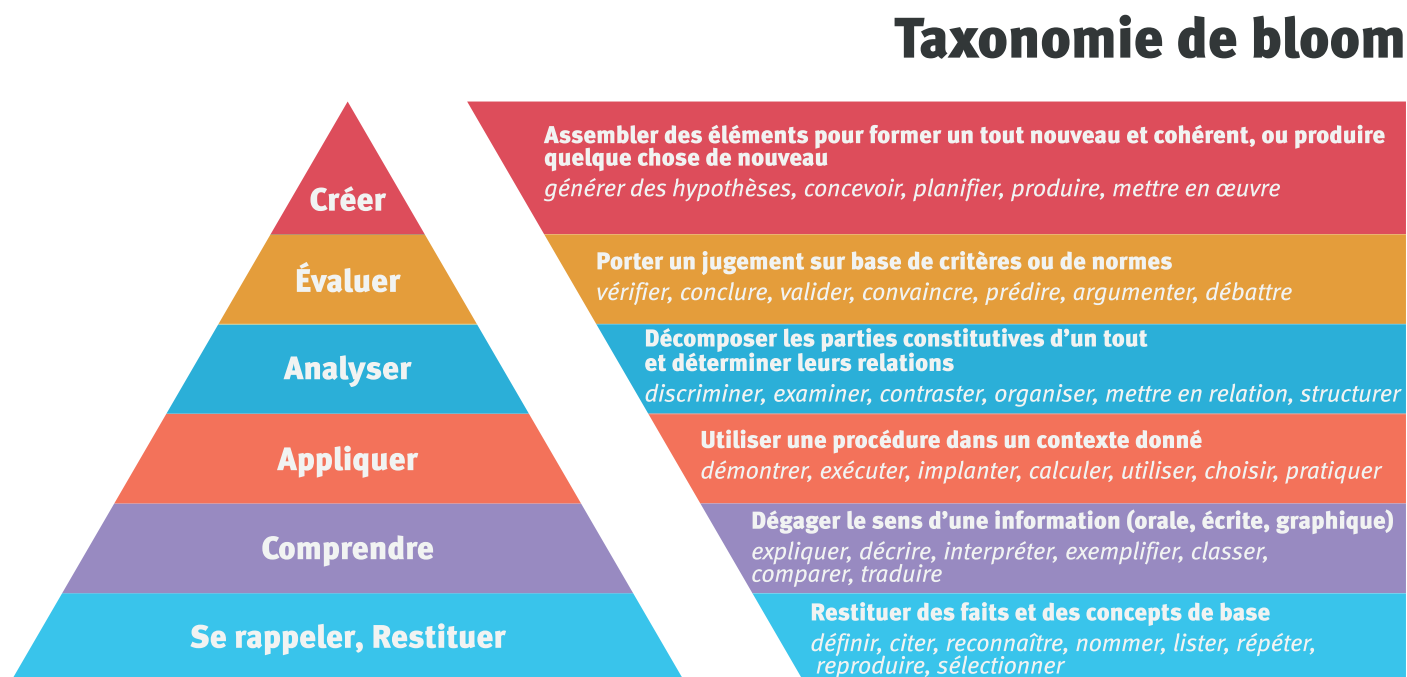
Taxonomie de bloom

Choisir les QCM, pour quel(s) niveau(x) de performance ?

Une autre étape importante, liée à la précédente, est la définition des niveaux de performance qui seront visés par l'évaluation.

La taxonomie de Bloom est sans doute la plus utilisée pour classifier les différents d'objectifs cognitifs. Révisée par Anderson et Krathwohl (2001), elle est composée de six catégories de performance allant du plus simple au plus complexe et chaque niveau requiert les capacités cognitives développées dans les niveaux inférieurs.

FIGURE 2. LA TAXONOMIE DE BLOOM, RÉVISÉE PAR ANDERSON ET KRATHWOHL (2001)



Selon la manière dont elles sont conçues, les QCM peuvent potentiellement évaluer les quatre premiers niveaux de la taxonomie de Bloom ; une QCM ne permet pas, en revanche, d'évaluer les niveaux « évaluer » et « créer » dans lesquels il est demandé à l'étudiant de *produire* une réponse plutôt que de la *reconnaître*. Pour tout ce qui concerne les capacités de production (ré-daction, expression orale, etc.), il est donc préférable d'utiliser d'autres modalités d'évaluation.

Niveaux	Exemples
Connaître : <i>L'étudiant doit pouvoir restituer et se rappeler de la définition telle qu'elle a été vue au cours</i>	Quelle est la définition d'un hôte paraténique ? - Hôte chez lequel le parasite se développe, mais sans atteindre sa maturité sexuelle - Hôte où ne s'effectue aucun développement du parasite, mais ce dernier demeure infectieux pour l'hôte définitif - Hôte chez lequel un parasite atteint sa maturité sexuelle et se reproduit - Hôte chez lequel un parasite peut se développer, atteindre sa maturité, mais ne peut se reproduire (Source : Guide d'élaboration d'un examen, CEFES, 2009)
Comprendre : <i>L'étudiant doit être capable d'associer un exemple à un concept théorique</i>	Quand Sarah a appris que les pingouins ne pouvaient pas voler, elle a modifié sa conception sur les oiseaux. Ceci illustre le processus de : - Accommodation - Conservation - Habitude - Égocentrisme (Source : Les cahiers du Louvain Learning Lab n°10, 2020)
Appliquer : <i>Pour résoudre le problème, l'étudiant doit appliquer une formule et trouver la réponse correcte</i>	Dans une population de chat, où les deux phénotypes «poils longs» (allèle c, récessif) et «allèle C, poils courts» (dominant) coexistent, on observe les fréquences génotypiques suivantes: $f(CC) = 49\%$, $f(Cc) = 47\%$, $f(cc) = 4\%$. Calculez la fréquence de l'allèle C et de l'allèle c. Parmi les réponses suivantes, quelle est la fréquence correcte ? $f(C) = 0,70$ $f(C) = 70$ $f(c) = 0,20$ $f(c) = 0,275$ $f(c) = 0,30$ - Toutes les réponses sont fausses (Source : Structures, fonctions, adaptations et diversité des animaux domestiques, Michaux, 2018)
Analyser : <i>L'étudiant doit pouvoir extraire les éléments clés du cas clinique avant de pouvoir poser le bon diagnostic</i>	Une femme de 24 ans, deux semaines après un accouchement, développe une céphalée, accompagnée de nausées et de vomissements, et fait une crise d'épilepsie tonico-clonique. A l'arrivée aux urgences, elle est en léger état confusionnel et un œdème papillaire est observé à l'examen du fond d'œil. Quel est le diagnostic le plus probable ? - Dissection carotidienne - AVC cardioembolique - Méningite - Thrombose de sinus veineux (Source : https://quizlet.com/be/352805364/neurologie-qcm-que-de-neurochir-neurophysio-euro-se-mio-flash-cards/)



QUELS TYPES DE QUESTIONS POUR DÉPASSER LA RESTITUTION ?

QCM et hauts niveaux d'acquisition

La capacité des QCM à évaluer des capacités supérieures à la restitution est souvent remise en question ; de nombreuses études ont tenté de répondre à cette critique en comparant les QCM avec d'autres modalités d'évaluation, tels que les travaux ou examens écrits qui exigent de l'étudiant une réponse construite. Plusieurs études (Becker & Johnston, 1999 ; Fredericksen, 1984, Hickson & Reed, 2011, Krieg & Uyar, 2011 in Scully, 2017) ont en effet mis en évidence la faiblesse des QCM dans l'évaluation de niveaux plus élevés d'acquisition, tandis que d'autres études (Hickson, Reed & Sander, 2012 ; Thissen, Wainer & Wang, 1994, Walstad & Becker, 1994 in Scully, 2017) n'ont pas réussi à démontrer une telle différence.

Selon Martinez (1999 in Scully, 2017), il est probable que les différences qui apparaissent soient moins le reflet des limites de l'évaluation par QCM que de l'utilisation qui en est faite. En effet, quand il est demandé à des experts de classer des QCM existantes en fonction des différents niveaux de la taxonomie de Bloom, il ressort que la plupart des items sont classés dans le niveau « restitution » (plutôt que dans les niveaux « compréhension », « application » ou « analyse »), ce qui suggère que les QCM pourraient potentiellement évaluer des niveaux d'acquisition plus complexes, mais que les items de type restitution sont souvent surreprésentés dans la pratique.

Il est également intéressant de noter que Bloom & Lavallée (1969) ne précisent pas où

se situe la limite entre les niveaux d'acquisition simples et complexes et les tentatives allant dans ce sens restent plutôt inconsistantes. C'est pourquoi Scully (2017) propose de concevoir la taxonomie de Bloom comme un continuum au sein duquel chaque niveau permet d'identifier un niveau supérieur à celui qui le précède.

Dans cette perspective, le terme « niveaux d'acquisition complexes » inclut tous les niveaux allant de la « compréhension » jusqu'à la « création », ce qui explique pourquoi nous parlerons souvent de questions visant à dépasser la restitution dans les lignes qui suivent.

L'enjeu est important : plusieurs études (Jensen et al., 2014 ; Leung, Mok & Wong, 2008 in Scully, 2017) ont montré que des étudiants qui expérimentent une évaluation mobilisant de hauts niveaux d'acquisition ont davantage tendance à adopter une approche en profondeur dans leur étude et ce, en opposition à un apprentissage en surface ou à du par cœur. Par conséquent, construire des QCM qui dépassent le niveau de la simple restitution peut encourager les étudiants à adopter des stratégies d'apprentissage en profondeur. On trouvera ci-après plusieurs suggestions pour construire des QCM visant des objectifs de plus haut niveau.

Utiliser des questions basées sur des situations concrètes

Une manière simple et efficace d'élaborer des QCM de qualité est de transformer des questions de restitution en question de niveaux cognitifs plus complexes. C'est notamment ce que Scully (2017) propose avec la notion de « flipped item » (ou item inversé). Par exemple, au lieu de proposer aux étudiants de reconnaître la définition correcte d'un concept parmi plusieurs propositions, il est possible de contextualiser la question en partant d'une situation concrète et en demandant d'identifier le concept illustré (voir exemple ci-dessous).

Transformer la question de la sorte permet de tester la compréhension ; la réponse ne pourra pas facilement être trouvée sur internet. Bien entendu, cette technique ne fonctionne que si l'exemple précis proposé n'a jamais été abordé au cours ou précédemment (sinon, il s'agirait de restitution).

TABLEAU 3. EXEMPLE DE « FLIPPED ITEM » OU « ITEM RENVERSÉ »

(Les cahiers du Louvain Learning Lab n°10, 2020) :

Item de base	Item renversé
Selon la théorie du développement cognitif de Piaget, qu'est-ce que l'accommodation ? a. La capacité à penser de façon logique b. La diminution d'une réponse à un stimulus fréquemment répété c. L'altération d'un schéma existant suite à une nouvelle information d. L'incapacité à comprendre des perspectives autres que les siennes	Quand Sarah a appris que les pingouins ne pouvaient pas voler, elle a modifié sa conception sur les oiseaux. Ceci illustre le processus de : a. Accommodation b. Conservation c. Habituation d. Égocentrisme



Utiliser des distracteurs de haute qualité

Quelle que soit la qualité de l'énoncé, il sera difficile d'évaluer le véritable degré de maîtrise des étudiants si les distracteurs ne sont pas plausibles.

Les distracteurs doivent idéalement exiger un haut niveau de discrimination pour pouvoir trouver la réponse correcte. Selon Morrison & Free (2001), « the student should be required to discriminate from among plausible alternatives. In other words, all choices provided should be relatively possible, but one choice should be better than the others ».

Typiquement, les questions qui exigent un calcul ou qui demandent d'identifier la solution la meilleure, la plus adaptée, la plus importante ou la plus prioritaire exigent un haut niveau de discrimination. Bien entendu, il faut veiller à ne pas introduire de subjectivité dans la proposition, puisque que la bonne réponse doit être incontestablement exacte et la seule correcte.

TABLEAU 4. QCM STANDARD ET QCM À HAUT NIVEAU DE DISCRIMINATION
(Morrison & Free, 2001, in Scully 2017) :

Énoncé et distracteurs standard	Énoncé et distracteurs avec haut niveau de discrimination
Which of the following assessment findings is characteristic of a client with Parkinson's disease? a. Night blindness b. Pain in lower extremities c. Shuffling gait d. Incontinence	A nurse is making a home visit to a 75-year-old male who has had Parkinson's disease for the past five years. Which of the following has the greatest implication for this client care? a. The client's wife tells the nurse that the grandchildren have visited for over a month b. The nurse notes that there are a numerous throw rugs throughout the client's home* c. The client has a towel wrapped around his neck that the wife uses to wipe her husband's face d. The client is sitting in an arms chair, and the nurse notes that he is gripping the arms of the chair

Dans l'exemple ci-dessus, l'item de droite exige un niveau supérieur de discrimination, car pour répondre correctement à la question, l'étudiant doit : (1) savoir que la démarche traînante est caractéristique de la maladie de Parkinson, (2) identifier le fait que les nombreuses carpettes de la maison du patient sont un potentiel danger pour ce dernier et (3) juger parmi toutes les propositions s'il s'agit bien de celle avec l'implication la plus importante pour le patient. Les trois distracteurs sont également plausibles ; C et D sont des symptômes de la maladie de Parkinson qui peuvent également avoir des implications pour le patient ; A peut potentiellement, mais pas nécessairement, avoir une implication (cela dépend du ressenti du patient face à la visite de ses petits-enfants).

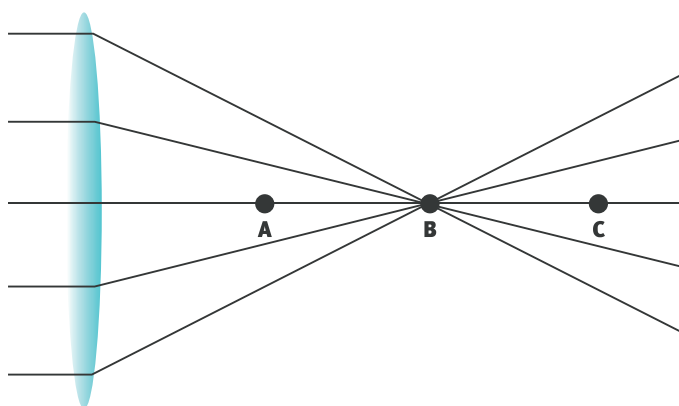
Une manière efficace de créer des distracteurs plausibles est d'anticiper les réponses erronées ou de directement s'inspirer des erreurs fréquentes des étudiants. C'est pourquoi il peut parfois être bénéfique de rédiger ses questions tout au long de l'année, plutôt que quelques jours avant l'examen. Lorsque

la correction est terminée, il est également recommandé d'analyser les réponses des étudiants à chaque question pour améliorer ou supprimer les distracteurs qui ne sont jamais sélectionnés par les étudiants (voir partie sur l'analyse et l'interprétation des résultats). En effet, il est préférable d'avoir trois leurres pertinents plutôt que cinq qui seraient éliminés très facilement, sans aucune maîtrise de la matière (Burton, 2001, 2004, Lesage et al., 2013 in Les cahiers du Louvain Learning Lab n°10, 2020).

Utiliser des éléments visuels à analyser

Il est également possible d'évaluer des niveaux supérieurs à la restitution en demandant aux étudiants d'analyser ou d'interpréter des éléments visuels (par ex. : graphiques, tableaux, images, vidéos, cartes, dessins, etc.) qui peuvent faire partie soit de l'énoncé, soit des propositions. L'utilisation d'éléments visuels implique généralement de plus hauts niveaux d'acquisition car l'étudiant doit appliquer un concept, analyser une situation concrète ou résoudre un problème pour répondre à la question. Ce type de questions peut également permettre de se rapprocher de certaines tâches issues du monde professionnel. Bien entendu, il est important de veiller à la qualité des images présentées pour éviter toute ambiguïté.

EXEMPLE DE QCM AVEC REPRÉSENTATION VISUELLE (Burton et Al, 1991) :



IN THIS DIAGRAM, PARALLEL LIGHT RAYS PASS THROUGH A CONVEX LENS AND CONVERGE TO A FOCUS. THEY CAN BE MADE PARALLEL AGAIN BY PLACING A:

- ☐ a. Concave lens at point B
- ☐ b. Concave lens at point C
- ☐ c. Second convex lens at point A
- ☐ d. Second convex lens at point B
- ☐ e. Second convex lens at point C*

sance de plusieurs concepts et la capacité à combiner ces derniers pour arriver à la réponse correcte. En d'autres mots, l'étudiant ne doit pas juste connaître les différents concepts, mais également comprendre la manière dont ceux-ci sont interconnectés.

Une façon simple de vérifier si un item répond à ce critère est de se demander si les étudiants doivent effectivement mobiliser plusieurs faits, concepts ou théories pour répondre. Les items favorisant la pensée multi-logique sont généralement (mais pas obligatoirement) contextualisés, c'est-à-dire intégrés dans des situations concrètes, comme vu précédemment. Dans l'exemple ci-dessous, l'étudiant doit connaître le médicament et ses contre-indications, mais il doit également être capable d'interpréter les différents résultats d'examen qui lui sont proposés.

EXEMPLE DE QCM FAVORISANT LA PENSÉE « MULTI-LOGIQUE » (Scully, 2017) :

WHICH OF THE FOLLOWING IS A CONTRAINDICATION FOR THE SPIRONOLACTONE?

- ☐ a. Serum creatinine = 3.0mg/dl*
- ☐ b. Serum potassium = 3.5mEq/L
- ☐ c. Resting heart rate = 68bpm
- ☐ d. Blood pressure = 130/85mmHg

Utiliser des questions de type « assertion-raison »

Une autre technique qui peut être utilisée pour évaluer des acquis de plus haut niveau est d'utiliser une forme de QCM appelée « question assertion-raison » ou « assertion-reason question (ARQ) » en anglais. Comme son nom l'indique, ce type de QCM comporte deux volets : une assertion (A) et une raison (R). L'étudiant doit d'abord décider de la véracité des deux propositions, pour s'interroger ensuite sur leur relation de causalité (la raison donnée est-elle bien une explication correcte de l'assertion ?).

Utiliser des questions qui favorisent la pensée « multi-logique »

Selon Morrisson & Free (2001), la pensée multi-logique dans l'élaboration de QCM (« multilogical thinking » ou « multiple-neuron items » en anglais) nécessite la connais-

EXEMPLE DE QUESTION DE TYPE « ASSERTION-RAISON » (Faillet, 2013) :

ASSERTION (A): LE FREE-MARTIN EST UN PHÉNOMÈNE NATUREL, OÙ DEUX FAUX JUMEAUX DE SEXE DIFFÉRENT SONT CONNECTÉS IN UTERO PAR VOIE SANGUINE, PAR L'INTERMÉDIAIRE DU PLACENTA. LA FEMELLE GÉNÉTIQUE (CARYOTYPE XX) NAÎT ALORS INTERSEXUÉE, C'EST-À-DIRE QU'ELLE EST MODIFIÉE DANS LE SENS MÂLE.

RAISON (R): LES TESTICULES DE L'EMBRYON BOVIDÉ MÂLE PRODUISENT UNE SUBSTANCE MASCULINISANTE, UNE HORMONE, VÉHICULÉE PAR LE SANG VERS L'EMBRYON FEMELLE.

- ☐ a. A et R sont vraies et R est l'explication correcte de A*
- ☐ b. A et R sont vraies et R n'est pas l'explication correcte de A
- ☐ c. A est vraie et R est fausse
- ☐ d. A est fausse et R est vraie
- ☐ e. A et R sont fausses

Pour chaque couple assertion-raison, il y a cinq réponses proposées dont une seule est correcte. Les questions assertion-raison sont particulièrement bien adaptées aux relations de cause à effet et elles sont plus exigeantes intellectuellement pour les étudiants que les QCM traditionnelles, car elles intègrent plusieurs concepts au sein d'une même question. Comme pour tout type de QCM, il faut être vigilant à ne pas trop complexifier linguistiquement la question, si l'on souhaite que celle-ci reste fiable et équitable. Comme ces questions sont également moins courantes que les QCM classiques, elles demandent au départ plus de préparation pour les enseignants et un peu d'entraînement de la part des étudiants.

Selon Bull & McKenna (2004, in Williams, 2006), il convient de faire attention à plusieurs points dans l'élaboration des questions de type assertion-raison :

- › La raison doit exister par elle-même pour pouvoir, dans un premier temps, être considérée séparément de l'assertion.
- › Les raisons dites « mineures » sont à éviter, car elles pourraient rendre la question ambiguë et faire l'objet de discussion.

- › Pour chaque item, il est préférable d'utiliser les cinq mêmes options (A-E).
- › Sur l'ensemble du questionnaire, la réponse correcte doit être distribuée équitablement à travers les cinq options.

En résumé, les principales pistes pour construire des QCM susceptibles d'évaluer des processus cognitifs dépassant la restitution sont les suivantes :

- › Transformer les questions de restitution en les intégrant dans des situations concrètes (ce qui aura en outre l'avantage de rendre la tricherie plus difficile si l'évaluation se fait en ligne).
- › Utiliser des distracteurs plausibles qui exigent un haut niveau de discrimination.
- › S'inspirer des erreurs fréquentes commises par les étudiants tout au long de l'année pour créer des distracteurs de qualité.
- › Analyser la manière dont sont distribuées les réponses des étudiants après la correction, pour améliorer ou supprimer certains distracteurs.
- › Intégrer le cas échéant des éléments visuels dans l'énoncé ou les solutions pour amener les étudiants à appliquer des concepts ou les analyser.
- › Formuler des questions exigeant de l'étudiant qu'il mobilise plusieurs concepts pour trouver la réponse correcte.
- › Utiliser des questions de type « assertion-raison » pour entraîner le raisonnement causal.



COMMENT RÉDIGER UNE QCM

La façon de rédiger les énoncés et les distracteurs influence grandement la qualité d'une QCM et la possibilité (ou non) pour l'étudiant de trouver la bonne réponse sans connaître la matière. On trouvera dans cette section quelques règles de base pour améliorer la formulation des consignes, des énoncés et des solutions associées.

La formulation des consignes

Comme le souligne Leclercq (1986), la consigne est d'une importance capitale et nécessite autant de soin que les autres composantes des QCM. La consigne est généralement valable pour toutes les questions d'une épreuve, ou pour toutes les questions de même type. Elle devrait toujours être écrite. Il est recommandé de la faire relire par un collègue ou un assistant pour vérifier sa clarté et sa compréhension, et de la communiquer aux étudiants en amont de l'épreuve.

La consigne précise idéalement l'ensemble des éléments suivants :

- › un rappel des acquis d'apprentissage qui sont mesurés par l'évaluation ;
- › le type de questions (par ex. : pour les questions qui suivent, une seule solution est considérée comme correcte) ;

- › le mode de réponse (par ex. : choisissez une seule des solutions proposées ou abstenez-vous de répondre) ;
- › les principes de notation (par ex. si vous trouvez la solution correcte, vous gagnez des points ; vous ne gagnez aucun point en cas de réponse incorrecte ou d'abstention) ;
- › le barème (par ex. : la notation se fera selon le barème suivant : réponse correcte = + 1 points, réponse incorrecte = 0 point, omission = 0 point) ;
- › la durée de l'épreuve (qu'il est possible d'estimer plus précisément en prétestant l'évaluation, et en gardant en tête qu'un expert est beaucoup plus rapide dans ses réponses).

Si l'épreuve est en ligne, il importe d'indiquer si l'étudiant peut naviguer comme il le désire dans le test (et éventuellement revenir sur des questions) ou pas (ces options correspondent aux modes de navigation « libre » ou « séquentielle » dans les paramètres d'un test Moodle). Lorsque le test est en ligne, il convient également de préciser si les étudiants peuvent poser des questions pendant l'épreuve et comment (par ex. : par Teams, par email, les deux...).

I La formulation de l'énoncé

Le tableau ci-dessous, ainsi que la plupart des illustrations, sont extraits ou adaptés de Bernard & Fontaine (1982), Bravard (2008), Leclercq (1986) et REPI (2005) in Lepage & Romainville (2009).

Règles de base	Contre-exemples
L'énoncé présente un seul problème à résoudre. Dans l'exemple ci-contre, les propositions a et b font référence à la définition de la presbytie, tandis que les propositions c et d ont trait à son traitement.	Un œil presbyte a. voit mal de loin b. voit mal de près c. doit toujours être corrigé d. ne doit être corrigé que pour la vision proche
L'énoncé adopte une formulation simple et claire qui contient tous les mots indispensables à sa compréhension. Dans l'exemple ci-contre, on ne sait pas si la question porte sur l'œuvre picturale, littéraire ou cinématographique.	Qui est l'auteur de « La jeune fille à la perle » ? a. Johannes Vermeer b. Tracy Chevalier c. Peter Webber
L'énoncé doit veiller à ne pas induire un jugement de valeur en évitant de faire appel à la subjectivité de l'étudiant ou à son pronostic personnel, contrairement à l'exemple ci-contre.	Lequel, parmi les peintres suivants, restera dans la postérité ? a. Picasso b. Renoir c. Van Gogh d. Magritte
L'énoncé sera formulé autant que possible, sous une forme positive et interrogative pour ne pas compliquer le traitement que l'étudiant doit opérer pour comprendre l'énoncé.	Aucun de ces gaz n'est un gaz organique, sauf un. Lequel ? a. dioxyde de chlore (ClO ₂) b. acide sulfhydrique (H ₂ S) c. ammoniac (NH ₃) d. acétone (CH ₃ COCH ₃) e. méthane (CH ₄)
Chaque question est indépendante des autres. Une question ne peut dépendre d'une réponse à une question précédente. Cela pénaliserait doublement un étudiant qui n'aurait pas trouvé la réponse à la première question	Q1. Quel est la valeur approchée du nombre π ? a. 3,141 592 653 589 793 b. 3,411 592 653 589 793 c. 3,194 872 653 589 793 Q.2 Quel est la circonférence d'un cercle dont le rayon est de 5 cm ? a. 31,9 cm b. 31,4 cm c. 34,1 cm

Afin de garantir la validité de l'épreuve, il est également important de ne pas trop complexifier la formulation des questions. En effet, le risque est alors de mesurer davantage la maîtrise de la langue plutôt que la connaissance de la matière ; c'est pourquoi il est préférable d'éviter les ambiguïtés et les tournures complexes comme les doubles négations.

I La formulation des solutions proposées

Le tableau ci-dessous, ainsi que la plupart des illustrations, sont extraits ou adaptés de Bernard & Fontaine (1982), Bravard (2008), Leclercq (1986) et REPI (2005) in Lepage & Romainville (2009).

Règles de base	Contre-exemples
La bonne réponse doit être incontestablement exacte et la seule correcte. Dans l'exemple ci-contre, les propositions a et b sont incontestablement fausses, mais la troisième pourrait être contestée, car elle donne la forme arrondie du nombre recherché.	Quelle est la valeur du nombre π ? a. 1,41 b. 2,74 c. 3,14
Les distracteurs doivent être incontestablement faux mais plausibles . Dans l'exemple ci-contre, les deux dernières propositions ne sont pas susceptibles de jouer leur rôle de distracteurs, puisqu'aucune d'entre elles n'est un indice de tendance centrale. De plus le terme « approprié » est assez ambigu.	Quel indice de tendance centrale serait le plus approprié pour la série 4, 3, 5, 4, 4, 6, 5,6, 3, 20 ? a. la médiane b. le mode c. la moyenne d. la variance e. l'écart-type
La bonne réponse ne doit être ni plus longue, ni plus explicite, ni plus élaborée que les autres solutions afin de ne pas fournir d'indice formel externe à la connaissance de la matière. Dans l'exemple ci-contre, la solution d s'impose du seul fait qu'elle est plus longue et plus technique	Qu'est-ce qu'un ostéopathe ? a. un insecte b. un coquillage c. un vêtement ancien d. une personne qui soigne par manipulation des os
La place de la bonne réponse parmi les solutions varie aléatoirement tout au long du test, mais lorsqu'il s'agit de noms propres ou de nombres, il est préférable de placer les solutions par ordre alphanumérique . Dans l'exemple ci-contre, la lecture est rendue plus difficile du fait que l'ordre numérique n'est pas respecté.	En quelle année, la loi belge accorde le droit de vote aux femmes (à l'exclusion des prostituées et des femmes adultères) pour les élections communales? a. 1938 b. 1934 c. 1950 d. 1948 e. 1920
Les réponses proposées doivent être homogènes dans leur contenu , leur forme et leur structure grammaticale , ce qui est une des règles les plus difficiles à respecter. Dans l'exemple ci-contre, les alternatives sont toutes en lien avec l'énoncé, mais les solutions b, c, e ne portent pas sur la même dimension (fonctions occupées par Lully) que a et d.	Jean-Baptiste Lully a. était chargé de la direction des « Petits violons » b. est mort d'un arrêt cardiaque c. est issu d'une famille de forgeron d. était le compositeur de Louis XIV e. menait ses musiciens à la baguette

Règles de base	Contre-exemples
Les solutions proposées ne doivent pas être des synonymes, s'exclurent mutuellement (car deux propositions qui se contredisent ne peuvent pas être toutes les deux correctes) ou s'imbriquer , contrairement à l'exemple ci-contre dans lequel les chevauchements sont multiples.	À quelle époque la bible de Gutenberg fut-elle imprimée ? a. entre 1415 et 1420 b. entre 1452 et 1454 c. entre 1450 et 1455 d. entre 1550 et 1555 e. au 15e siècle
Pour éviter qu'un mot ne se répète dans chacune des solutions proposées, on intègre d'habitude tout ce qui est commun aux réponses dans l'énoncé . Dans l'exemple ci-contre, il aurait été plus adéquat de poser la question en ces termes : « la molysmologie est la science des... »	Qu'est-ce que la molysmologie ? a. science des microbes b. science des molécules c. science des mollusques d. science des pollutions e. science des singes
Il est préférable d'éviter que certains mots de l'énoncé ne se répètent dans une des solutions ou qu'une grande proximité d'expressions contenues dans la question et dans une des solutions ne suggère trop facilement la bonne réponse.	Qu'est-ce qu'un patriclan ? a. un système politique isolationniste b. une réunion annuelle des hommes dans une société primitive c. un clan fondé sur la filiation patrilinéaire

Quelques conseils supplémentaires au sujet de la présentation des solutions proposées :

- la position de la bonne réponse parmi les solutions proposées doit varier aléatoirement tout au long du test de manière à être distribuée plus ou moins équitablement ;
- le fait de positionner plus de réponses correctes en C ou D ne permet pas de « mieux les cacher », car ce sont ces propositions qui sont choisies préférentiellement par les étudiants qui répondent au hasard (Les cahiers du Louvain Learning Lab n°10, 2020) ;
- pour réduire l'effet du hasard, utiliser au minimum trois propositions de réponse pour chaque question, tout en veillant à ce que les distracteurs restent plausibles (raison pour laquelle les Vrai-Faux sont généralement déconseillés, sauf cas spécifiques).





BARÈME(S) DE CORRECTION ET CORRECTION DU HASARD

Selon Leclercq (1986), un barème est « *un ensemble de tarifs, le tarif étant une conséquence convenue à l'avance pour un type de réponse précis* ». Le choix du barème, ainsi que la pondération éventuellement utilisée, doivent être définis avant la correction de l'épreuve pour que les scores calculés prennent ceux-ci en considération. Il faut évidemment que l'étudiant soit informé avant de commencer l'épreuve des conséquences, en termes de points, de ses choix de réponse(s) ou de son abstention.

Il existe de nombreux barèmes possibles, nous en aborderons ci-après quelques-uns.

Le barème classique (sans pénalisation)

Dans le barème sans pénalisation, l'enseignant attribue un point en cas de bonne réponse et zéro en cas d'omission ou de réponse incorrecte.

L'avantage du barème classique est sa simplicité et sa transparence.

Au rang des inconvénients, le barème sans pénalisation ne prend pas en compte l'effet dû à une réponse donnée au hasard. Dans le cas par exemple d'un questionnaire comprenant vingt QCM à quatre solutions, un étudiant qui répondrait totalement au hasard a une chance sur quatre de répondre correctement à chaque question et obtiendrait donc en moyenne un score de 5/20 ; en étant chanceux, il peut évidemment obtenir davantage. Le barème classique ne valorise pas non plus le fait que l'étudiant puisse reconnaître son ignorance.

Ci-après, nous aborderons diverses méthodes pour tenter d'atténuer les effets du hasard.

I Le barème utilisant la correction pour devinette

Le recours à des points négatifs est souvent utilisé pour réduire l'effet du hasard – c'est ce que l'on appelle la **correction pour devinette** (*correction for guessing en anglais*).

Le principe est d'attribuer des points aux réponses correctes, de ne pas pénaliser les omissions, et de retirer des points en cas de réponse incorrecte, afin de pousser l'étudiant à se montrer prudent s'il ne connaît pas la réponse. Le nombre de points retranchés dépend du nombre de solutions proposées (k), selon les formules suivantes :

BARÈMES		TARIFS		
		Réponse correcte	Omission	Réponse incorrecte
avec pénalisation	1 solution correcte (QRU)	+1	0	$-1/(k-1)$
	c solutions correctes (QRM)	+1	0	$-c/(k-c)$

Ce barème est régulièrement critiqué pour diverses raisons (pour un résumé, voir Uyttebrouck, 2014). La principale critique porte sur le fait que la correction pour devinette, contrairement à ce que l'on pourrait penser, pénalise les étudiants prudents ou peu confiants qui ne sont pas encouragés à faire la démonstration de leurs connaissances par crainte de perdre des points, tandis qu'elle favorise les étudiants les plus téméraires qui tentent leur chance. Pour les étudiants, on peut ainsi montrer que la meilleure stratégie consiste à tenter de répondre à toutes les questions, alors même que les consignes de l'enseignant affirment le contraire : l'étudiant est donc induit en erreur sur la meilleure stratégie à adopter. Un biais de genre en faveur des garçons a également été mis en évidence dans certaines recherches

(Siegfried, 1979; Buck & al., 2002; Leaver and van Walbeek, 2006 in Du Plessis & Du Plessis, 2009), car ils seraient plus enclins à prendre des risques. Par ailleurs, il s'agit d'une méthode qui donne une trop grande place à des considérations stratégiques sans lien avec la maîtrise de la matière.



En raison de ces critiques, l'ULB, comme d'autres universités, en déconseille désormais l'utilisation (l'Université de Gand, par exemple, a interdit la correction par devinette dès 2013).

Le barème relevant le score minimal à atteindre

Plusieurs universités (l'Université de Gand, l'UCL, ...) proposent, pour remplacer la correction par devinette, la méthode dite du **standard setting**. Le principe est ici de contrer les effets du hasard en **augmentant le nombre de bonnes réponses nécessaires** : l'étudiant doit répondre correctement à plus de la moitié des questions pour réussir le test.

La distinction entre score et note

Avant d'aller plus loin, il est essentiel de rappeler la différence entre score et note, qui correspondent à deux étapes différentes du processus d'évaluation : la mesure « brute » de la performance d'abord, puis l'attribution à cette performance d'une note (sur 20 chez nous) qui décide du degré de réussite. « En anglais cette distinction s'appelle scoring pour la première étape et grading pour la seconde. Or, ces deux étapes sont trop souvent confondues » (Rege Colet & Sylvestre, 2013, p. 287).

Dans certains cas, il peut arriver que l'échelle de mesure et l'échelle de notation aient la même forme. Par exemple, dans un questionnaire comportant 20 questions, notées 1 point chacune, le score est calculé sur 20, et la note attribuée sera généralement identique au score. Ce choix, qui vise généralement à simplifier les calculs, n'est pas problématique en soi, à condition que la similarité des échelles ne fasse pas oublier qu'il y a bien deux échelles différentes et que l'équivalence posée entre les deux échelles résulte d'un choix. Pour bien faire la différence, il suffit de rappeler par exemple qu'en Suisse, la note serait exprimée sur 6, non sur 20. Aux USA, l'échelle de notation utiliserait les lettres A, B, C, D et F ; l'enseignant américain recourant au même questionnaire de 20 questions devra décider quel score correspond à quelle note, et ce qui constitue une performance minimale pour réussir (Uyttebrouck, 2014).

Dans le cas du standard setting, le **score** minimal à atteindre sera légèrement supérieur à la moitié du score maximal possible afin de compenser l'effet du hasard, même si l'étudiant devra toujours obtenir une **note** de 10 pour réussir.

Calcul du score minimal à atteindre

La formule de calcul du **seuil** ou **score minimal à atteindre** pour réussir, pour N questions à n solutions proposées, est la suivante :

$$\text{Score minimal} = (N + N/n) / 2$$

Dit autrement, au lieu de placer la barre à la moitié de réponses correctes comme on le fait généralement, on ajoute la moitié des points qui seraient obtenus si l'étudiant répondait totalement au hasard. Dans le cas d'un QCM à 4 propositions de réponses par exemple, répondre au hasard permettrait statistiquement d'obtenir 5/20. La barre sera donc mise à $(20 + 5) / 2$ soit 12,5.

Les seuils les plus fréquents sont ainsi :

Nbre de propositions	2	3	4	5
Seuil en %	75 %	66,67 %	62,50 %	60 %
Seuil sur 20	15	13,33	12,5	12

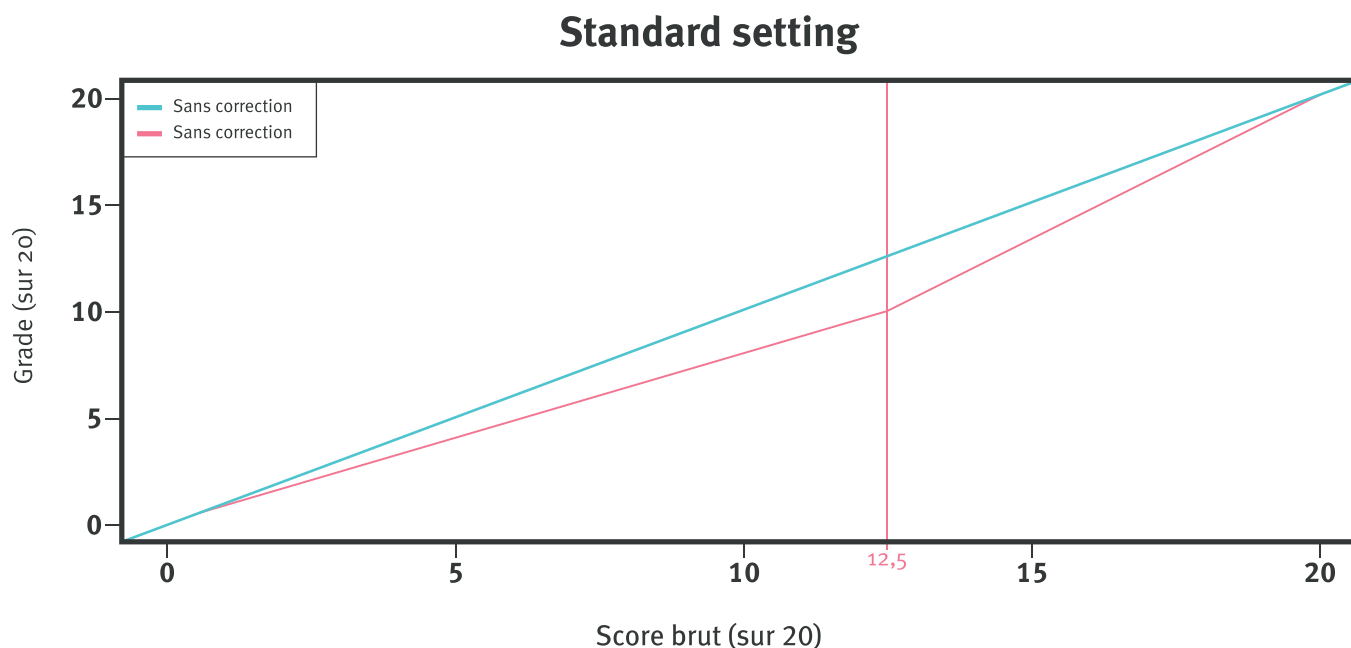
Le service Support en Méthodologie et Calcul Statistique (SMCS) de l'UCLouvain propose un [outil en ligne](#) permettant aisément de calculer le seuil, notamment pour les épreuves mêlant plusieurs types de questions avec des nombres différents de propositions.

Calculer la note de l'étudiant

Différentes formules existent. Nous proposons ici également de retenir celle adoptée par l'UCLouvain qui fournit également un fichier Excel permettant de transformer aisément les scores des étudiants en notes.

Sans entrer dans le détail des formules (disponibles au besoin sur le site ci-dessus), le principe est de transformer les scores de telle sorte qu'un score de 0 donne une note de zéro, un score maximum donne une note de 20, et un score égal au seuil donne une note de 10.

Visuellement, voici un graphe montrant la différence entre barème classique et standard setting, pour une épreuve de 20 questions à 4 propositions de réponse :



Source : graphique généré par l'application du SMCS de l'UCLouvain

Avantages et inconvénients

Numériquement, le barème relevant le score minimal n'est pas « plus généreux » que la correction pour devinette, au contraire : le résultat est identique à ce que l'on obtient via la correction par devinette où l'étudiant n'omet aucune réponse. **La différence majeure est que l'on n'a pas induit l'étudiant en erreur sur la meilleure stratégie à adopter** : il est dans l'intérêt de l'étudiant de répondre à toutes les questions, même celles pour lesquelles il hésite. L'étudiant peut se concentrer uniquement sur la démonstration de sa connaissance et laisser les considérations stratégiques de côté. **L'évaluation est plus équitable** car elle ne favorise plus certains profils d'étudiants par rapport à d'autres.

Comme tout barème, le standard setting n'est cependant pas parfait, on le voit : il est plus complexe à calculer que le barème classique, il incite l'étudiant à répondre même s'il ne connaît pas la réponse, et il demande de bien expliquer aux étudiants la différence entre score et note.

Le barème utilisant les degrés de certitude

Pour rappel, dans ce type de QCM, l'étudiant doit d'abord sélectionner la réponse qui lui semble correcte parmi les propositions, puis doit y associer un degré de certitude. Il existe un grand nombre d'échelles de certitude ; avec des pourcentages, avec des intervalles symétriques, asymétriques ou qui varient en fonction du nombre d'intervalles. Ces différences sont principalement méthodologiques et nous ne rentrerons pas ici dans les débats docimologiques sous-jacents. L'échelle que nous proposons de retenir est celle de Leclercq et Poumay (2003) reprise dans la consigne suivante : **« Fournissez comme degré de certitude un des six pourcentages de chances suivants : 0%, 20%, 40%, 60%, 80%, 100% ».**

Le barème des tarifs lié aux degrés de certitude a également fait l'objet de nombreuses recherches et de longs débats. Selon Leclercq et Poumay (2003), les évaluations utilisant des degrés de certitude doivent avant tout assurer une certaine validité conséquentielle, ce qui signifie qu'elles doivent être capable d'entraîner une réflexion métacognitive et des régulations sur les stratégies de réponses et d'apprentissage. Pour cela, il faut, d'une part, que les indices fournis aux étudiants sur leur métacognition aient une certaine valeur diagnostique en précisant ce qui doit être amélioré et, d'autre part, que la métacognition soit évaluée le plus pertinemment possible et que son poids dans le score total à une épreuve soit ressenti comme juste par les étudiants et les enseignants.

La solution que proposent les auteurs pour répondre au critère de validité conséquentielle est d'utiliser trois indices de réalisme dans l'auto-évaluation des performances :

- › L'indice de confiance : la moyenne des degrés de certitude accompagnant les réponses correctes. Idéalement, elle devrait être la plus élevée possible (proche de 100%).
- › L'indice de prudence : la moyenne des degrés de certitude accompagnant les réponses incorrectes. Idéalement, elle devrait être la plus faible possible (proche de 0%).
- › L'indice de discriminance (ou nuance) : la différence entre l'indice de confiance et de prudence.

Sur cette base, Leclercq et Poumay (2003) proposent d'accorder aux étudiants 3 points de bonus potentiels sur 20 sur la base des critères suivants :

- › + 1 point si l'indice de confiance est supérieur à 50% ;
- › +1 point si l'indice de prudence est inférieur à 50% ;
- › +1 point si l'indice de discriminance est supérieur 20%.

On notera que, dans ce système, les informations tirées des degrés de certitude ne peuvent faire que gagner des points aux étudiants, jamais en perdre.

A titre d'illustration, prenons le cas de quatre étudiants. Le premier se montre relativement confiant quand il sait, mais pêche par manque de prudence : il reçoit un bonus d'un point de confiance. Le deuxième affiche le profil inverse : il reçoit un bonus d'un point de prudence. Le troisième s'autoévalue correctement, mais reste relativement confiant ou prudent face à ses réponses ; il obtient deux points de bonus métacognitif. Le quatrième étudiant est à la fois confiant dans ses réponses quand il a raison, prudent quand il se trompe, avec en outre un indice de discriminance supérieur à 20% - il obtient 3 points de bonus.

	Confiance		Prudence		Discriminance		Score classique	Bonus
Étudiant 1	72%	+1	57%	0	19%	0	8	9
Étudiant 2	48%	0	45%	+1	3%	+1	8	9
Étudiant 3	58%	+1	40%	+1	18%	+1	8	10
Étudiant 4	72%	+1	35%	+1	37%	+1	8	11

Le barème proposé par Leclercq et Poumay (2003) est basé sur la théorie des décisions, selon laquelle la seule stratégie possible pour maximiser le score est celle de répondre le plus honnêtement possible. Autrement dit, les étudiants qui s'auto-évaluent correctement, c'est-à-dire sans trop de sur ou sous-estimation, récoltent le plus de points.

Autres pistes pour diminuer les effets du hasard

Outre le fait de soigner la construction et la rédaction des QCM pour ne pas donner d'indices concernant la bonne réponse et utiliser des distracteurs de qualité, la solution la plus simple pour diminuer la part de hasard consiste à augmenter le nombre de solutions proposées, tout en veillant à ce que les distracteurs restent plausibles. Cette approche a l'avantage d'être compatible avec le barème classique : avec 6 propositions de réponse au lieu de 4 par exemple, le gain théorique apporté par les réponses au hasard passe de 0.25 points par question (en supposant que la question vaille 1) à 0.166.

Les solutions générales (voir supra) sont également une manière d'ajouter assez facilement des propositions.

Il est également important de proposer un nombre de questions suffisant (sans transformer pour autant l'épreuve en marathon bien sûr) : les chances d'obtenir un bon résultat grâce au seul hasard diminuent avec le nombre de questions.

A titre d'exemple, voici les chances d'obtenir une note de 8/20 ou plus en répondant totalement au hasard dans différentes configurations :

- Avec 20 questions et 4 choix : 4%
- Avec 20 questions et 5 choix : 1%
- Avec 30 questions et 4 choix : 2%
- Avec 30 questions et 5 choix : 0.3%
- Avec 40 questions et 4 choix : 1.2%
- Avec 40 questions et 5 choix : 0.1%

Enfin, rappelons que les QCM ne sont pas les seules formes de questions possibles et qu'il est conseillé, quand la situation le permet, d'inclure également quelques questions ouvertes dans l'épreuve.



COMMENT ANALYSER ET INTERPRÉTER LES RÉSULTATS ?

Savoir comment interpréter et utiliser les informations relatives aux résultats des étudiants est aussi important que le fait de savoir comment construire un questionnaire de qualité (Crisp et Palmer, 2007). Grâce à ces informations, il est possible :

- › de se faire une idée générale du niveau de connaissance des étudiants ;
- › de prendre des décisions au sujet d'items éventuellement problématiques ;
- › de modifier l'ordre des questions, si nécessaire, en fonction de leur degré de difficulté ;
- › d'améliorer la qualité de chaque item et plus précisément des distracteurs qui ne sont jamais choisis.

L'analyse des résultats constitue ainsi une partie essentielle du processus d'évaluation par QCM qui repose sur différents indicateurs

statistiques :

- › l'indicateur statistique p' qui prend en compte la difficulté des questions ;
- › l'indicateur statistique $rpbis$ qui prend en compte le pouvoir discriminant des questions ;
- › l'analyse de la qualité des distracteurs utilisés.

Notons néanmoins que ce type d'analyse ne peut se faire que pour un nombre suffisant d'étudiants. Selon Dessus (2015), tout calcul réalisé sur des échantillons de moins de 20 étudiants est à prendre avec la plus grande précaution. Ces indicateurs sont également dépendants du niveau des étudiants. Autrement dit, il est probable qu'un item jugé comme très difficile pour un groupe ne le soit pas forcément pour un autre groupe. Dessus (2015) suggère que dans les deux cas, le jugement doit prévaloir.

La prise en compte de la difficulté des questions

Le degré de difficulté de chaque question peut se mesurer à l'aide de l'indicateur statistique p' qui est la proportion d'étudiants répondant correctement à une question. Il se calcule en divisant le nombre d'étudiants qui ont répondu correctement à la question par le nombre total d'étudiants qui ont répondu à la question. Sa valeur est comprise entre 0 et 1. Plus la valeur de p' est proche de 1, plus la question est considérée comme « facile ».

Selon McAlpine (2002b, in Dessus, 2015), si la valeur de p' est supérieure à 0.85, cela signifie que la question est trop facile. Inversement, si la valeur de p' est inférieure à 0.15, cela signifie que la question est trop difficile. Dans les deux cas, McAlpine suggère d'écarter ces questions. Ces valeurs limites varient légèrement en fonction des auteurs qui se sont penchés sur la question. De plus, des raisons pédagogiques peuvent amener à garder des questions au-delà ou en-deçà de ces seuils.

L'indicateur statistique p' n'est donc pas suffisant en lui-même pour juger de la qualité des items. Pour décider s'il convient de supprimer ou non un item, il est nécessaire de s'intéresser au pouvoir discriminant de chaque question.

La prise en compte du pouvoir discriminant des items

Le deuxième indice utilisable concerne le pouvoir discriminant de chaque item. Il permet de déterminer si une question discrimine correctement les étudiants performants et les étudiants les plus faibles. L'indicateur statistique rpbis, qui est une variante du coefficient de corrélation de Pearson (utilisée lorsque l'une des deux variables est dichotomique), est nommé « coefficient de corrélation bisérial de point » (rpbis). Il repose sur le calcul des corrélations entre le score obtenu à chaque item et le score total des autres questions.

Comme tout coefficient de corrélation, sa valeur est comprise entre -1 et +1.

Plus précisément, la valeur du rpbis à un item donné peut être :

- › Supérieure à 0,20 (question discriminante) : ce qui signifie que les étudiants qui réussissent bien à l'ensemble du test réussissent également bien à l'item concerné, tandis que les étudiants peu performants à l'ensemble du test réussissent moins bien à l'item concerné.
- › Inférieure à 0,20 (question non discriminante) : ce qui signifie que les étudiants qui performant peu à l'ensemble du test réussissent mieux à l'item concerné que les étudiants qui réussissent bien à l'ensemble du test. Ce cas de figure peut laisser penser que cet item est un faible prédicteur de la connaissance ou compétence mesurée dans le test et il convient dès lors d'envisager sa suppression.

Pour les tests élaborés via l'Université Virtuelle (Moodle), les indices de facilité et de discrimination sont automatiquement calculés et disponibles (voir *Administration du test* › *Résultats* › *Statistiques*).

L'analyse de la qualité des distracteurs utilisés

En plus des indices de difficulté ou de discrimination, on peut également analyser item par item la fréquence des choix des distracteurs par les étudiants. Il convient en effet de vérifier si les distracteurs ont bien joué leur rôle, en évitant qu'un distracteur soit sélectionné par tout le monde ou inversement par personne. Selon Johnstone (2003 in Dessus, 2015), un distracteur joue son rôle si environ 5% au moins des personnes le choisissent, mais ce seuil est à utiliser avec prudence, vu que la valeur de choix maximale dépend bien sûr du nombre de propositions de réponses.

Une manière simple et efficace de déterminer si les distracteurs jouent bien leur rôle est d'utiliser un tableau de fréquence ou de représenter graphiquement une distribution par item et par réponse

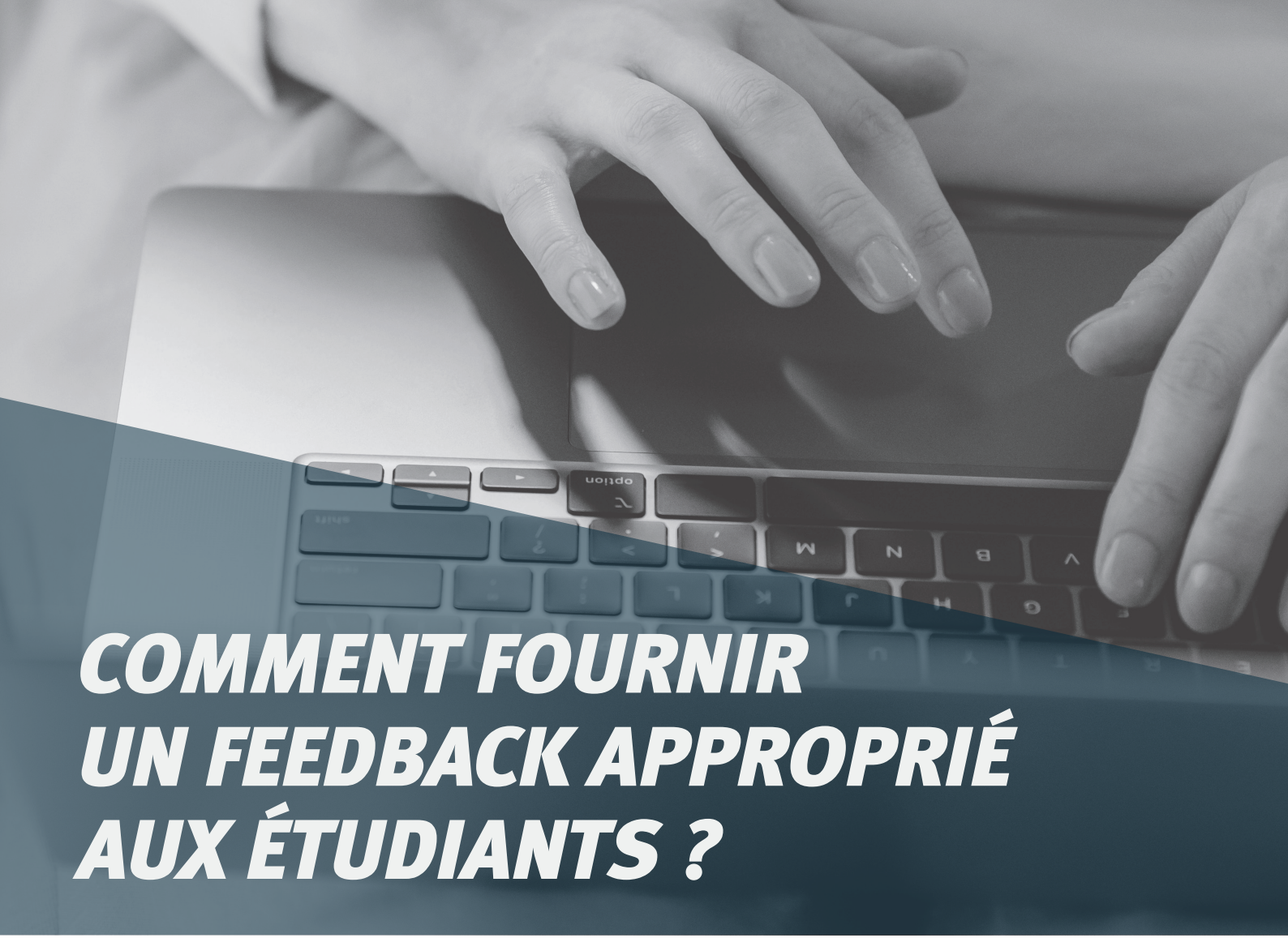
(à nouveau, le tableau des fréquences est disponibles dans les statistiques du test dans Moodle). Il est ainsi possible de vérifier visuellement la distribution des choix. Si un distracteur n'est que très rarement ou jamais choisi, il doit être supprimé ou modifié. Par contre, si un distracteur est choisi plus fréquemment ou tout autant que la réponse correcte, il s'agit probablement d'un distracteur qui prête à confusion.

I Attention à la constante macabre !

Une mise en garde s'impose quant au bon usage de l'analyse d'items. On appelle « effet Posthumus » (du nom d'un chercheur hollandais, K. Posthumus) la tendance des enseignants, consciente ou inconsciente, à noter les apprenants en respectant une courbe de Gauss plus ou moins centrée sur la moyenne (10/20), et ceci quelle que soit la maîtrise du cours par les apprenants. En d'autres termes, si les étudiants réussissent mieux, l'enseignant peut avoir tendance à augmenter le nombre ou la difficulté des questions de façon à toujours obtenir une proportion similaire de réussites et d'échecs. C'est cette proportion récurrente d'échecs, quelles que soient les qualités de la cohorte, qu'André Antibé (2003) appelle la constante macabre - Philippe Meirieu parle lui de « dictature de la courbe de Gauss ».

L'analyse d'items et les décisions de régulation qui en résultent doivent avoir pour but d'améliorer la fidélité des épreuves, et non de s'assurer d'un nombre stable d'échecs ou de réussites. A titre d'exemple, on peut imaginer une question portant sur un objectif essentiel de la matière, sur lequel l'enseignant a particulièrement insisté et pour lequel les étudiants se sont bien préparés : si une majorité d'étudiants répondent correctement, la valeur élevée de p' (indice de facilité) n'est pas un problème mais plutôt une « bonne nouvelle » prouvant que l'enseignant a atteint ses objectifs ; retirer cette question portant sur un objectif jugé central ne serait pas forcément judicieux.





COMMENT FOURNIR UN FEEDBACK APPROPRIÉ AUX ÉTUDIANTS ?

Afin de boucler le processus d'évaluation, il est important de ne pas oublier de fournir un feedback constructif aux étudiants, c'est-à-dire un feedback qui leur permettra de s'améliorer. Pour rappel, le feedback est « l'information que l'enseignant fournit à l'étudiant à propos de la réalisation des tâches d'apprentissages » (Nicol & MacFarlane-Dick, 2006). Il comprend le retour sur la performance de l'étudiant, mais également ce qu'il doit faire pour s'améliorer. Il peut se concevoir en cours d'année, dans le cadre d'une évaluation formative, ou à l'issue de l'évaluation des acquis dans le cadre d'une évaluation sommative.

Nicol & MacFarlane-Dick (2006) distinguent sept caractéristiques d'un feedback de qualité :

- › il aide à clarifier ce qu'est une bonne performance (objectifs, critères et attendus) ;
- › il offre la possibilité de combler l'écart entre la performance actuelle et celle

souhaitée ;

- › il fournit des renseignements de qualité sur l'apprentissage ;
- › il encourage le dialogue autour de l'apprentissage ;
- › il encourage la motivation et l'estime de soi ;
- › il facilite le développement de l'auto-évaluation (réflexion) dans l'apprentissage ;
- › il fournit à l'enseignant des informations qui peuvent être utilisées pour aider à façonner l'enseignement.

En ce qui concerne plus précisément la visite des copies, les conditions dans lesquelles les étudiants peuvent consulter leurs épreuves sont définies par le [Décret Paysage](#) et dans l'article 89 du [Règlement général des études](#) de l'ULB.

RESSOURCES BIBLIOGRAPHIQUES

- › ANDERSON, LW. (Ed.), KRATHWOHL, D.R. (Eds.). (2001). A taxonomy for learning, teaching and assessing: A revision of Bloom's Taxonomy of Educational Objectives. New-York: Longman.
- › BÉLANGER, C. (2009). Guide d'élaboration d'un examen. Centre d'études et de formation en enseignement supérieur (CEFES), Université de Montréal. En ligne (consulté le 15/03/21) :
<https://bit.ly/3xWzTr1>
- › BETTS, L.R., ELDER, T.J., HARTLEY, J. & TRUEMAN, M. (2009). Does correction for guessing reduce students' performance on multiple-choice examinations? Yes? No? Sometimes? Assessment & Evaluation in Higher Education, 34 (1), 1-15.
- › BLOOM, B., & LAVALLEE, M. (1969). Taxonomie des objectifs pédagogiques. Montréal, CANADA : Éducation Nouvelle.
- › BOUVY, T., WARNIER, L. (2011). Document de synthèse sur les QCM, IPM, document interne.
- › BURTON, S.J., SUDWEEKS, R.R., MERRILL, P.F., & WOOD, B. (1991). How to Prepare Better Multiple-Choice Test Items: Guidelines for University Faculty. Brigham Young University Testing Service and the Department of Instructional Science. En ligne (consulté le 15/03/21) :
<https://bit.ly/3teliuk>
- › BUSH, M. (2001). A Multiple-Choice Test that Rewards Partial Knowledge. Journal of further and Higher Education, 25(2), 157-163. En ligne (consulté le 15/02/2021):
<https://bit.ly/3tkSCKu>
- › BUSH, M. (2014). Reducing the need for guesswork in multiple-choice tests, Assessment & Evaluation in Higher Education, 40 (2), pp. 218-231. En ligne (consulté le 15/02/2021):
<https://bit.ly/3b3lt6j>
- › CONNIE, M. (2012). Writing Multiple Choice Questions for Higher Order Thinking. En ligne (consulté le 08/02/2021):
<https://bit.ly/3tkaGuY>
- › CRISP, G.T., & PALMER, E.J. (2007). Engaging academics with a simplified analysis of their multiple-choice question (MCQ) assessment results. Journal of University Teaching & Learning Practice, 4(2).

- › DEANE, F., & BOZIN, D. (2017). Using Guiding Principles to Construct Effective Multiple-Choice Exams to Assess Legal Reasoning. *Legal Education Review*, 26(1).
- › DESSUS, PH. (2015). Tutoriel : Analyser les items d'un questionnaire. Université Grenoble Alpes. En ligne (consulté le 26/03/2021):
 <https://bit.ly/3eRv9IS>
- › DICKINSON, M. (2011). Writing Multiple Choice Questions for Higher Order Thinking. *Learning Solutions Magazine*. En ligne (consulté le 08/02/2021):
 <https://bit.ly/33gRT9o>
- › DU PLESSIS, S. & DU PLESSIS, S. (2009). A new and direct test of the 'gender biaisi' in multiple-choice questions. *Working Papers 23/2009*, Stellenbosch University, Department of Economics.
- › FELLESEN, M. (2004). Using assessment to support higher level learning: the multiple-choice item development assignment, *Assessment & Evaluation in Higher Education*, 29 (6), 703-719.
- › Ghent University (s.d). No more negative marking in multiple choice questions.
 <https://bit.ly/3xL7fRv>
- › HOWE, R. (2018). Des QCM pour évaluer des compétences. Et pourquoi pas ? Portail du réseau collégial du Québec. En ligne (consulté le 08/02/2021):
 <https://bit.ly/33ixGAa>
- › KREBS R. (2004). Mode d'emploi pour la création de questions et d'exams à choix multiples pour l'enseignement de la médecine, Berne : Institut für Medizinische Lehre, Abteilung für Assessment und Evaluation.
- › LECLERCQ, D. (1986). La conception des questions à choix multiple, Bruxelles : Labor.
- › LECLERCQ, D., & POUMAY, M. (2003). Trois nouveaux indices de réalisme dans l'auto-évaluation des performances, *Cahiers du Service de Pédagogie expérimentale*, Université de Liège.
- › LECLERCQ, D. (2006) L'évolution des QCM. In G. Figari et L. Mottier-Lopez (dir.), *Recherches sur l'évaluation en Education*. Paris : L'Harmattan, 139-146.
- › LECLERCQ, D., & POUMAY, M. (2007). Comment savoir ce que l'on sait ? Dans D. Leclercq. *Psychologie Éducationnelle de l'Adolescent et du jeune Adulte*. Liège : Éditions de l'Université de Liège.
- › LEPAGE, P. & ROMAINVILLE, M. (2009). RESEAU 69 : Le Questionnaire à Choix Multiple. Service de Pédagogie Universitaire.
- › MALAU-ADULI, B.S., ZIMITAT, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, 37(8), 919-931.
- › MCCOUBRIE, P. (2004). Improving the fairness of multiple-choice questions: a literature review, *Medical Teacher*, 26(8), 709-712.
- › MORRISON, S. & FREE, K. (2001). Writing multiple-choice test items that promote and measure critical thinking. *Journal of Nursing Education* 40: 17-24. En ligne (consulté le 08/02/2021):
 <https://bit.ly/3h255Hb>



- › NICOL, D.J., & MACFARLANE-DICK, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218.
- › PELLEGRINO, J. W., CHUDOWSKY, N., & GLASER, R. (Eds.). (2001). *Knowing what students know : The science and design of educational assessment*. Washington, DC : National Academy Press.
- › RAUCENT, B. & VANDER BORGHT C. (2006). Chapitre 4. Comment évaluer les acquis des étudiants : former ? certifier ? Dans B. Raucent (dir.). *Être enseignant: Magister ? Metteur en scène ?* Louvain-la-Neuve, Belgique: De Boeck Supérieur, 237-299.
- › SCULLY, D. (2017). Constructing Multiple-Choice Items to Measure Higher-Order Thinking. *Practical Assessment, Research and Evaluation*: Vol. 22, Article 4. En ligne (consulté le 08/03/2021):
<https://bit.ly/2Rri1vw>
- › 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).
- › Université Catholique de Louvain, Les cahiers du Louvain Learning Lab (2016). *Carnet de l'enseignant. Voyages en pédagogie universitaire*. En ligne (consulté le 18/09/17) :
<https://bit.ly/3hfZ5uP>
- › Université Catholique de Louvain, Les cahiers du Louvain Learning Lab N°10 (2020). QCM or not QCM ? Processus de conception d'une évaluation par QCM. Sous la direction d'Émilie Malcou rant. Presses Universitaires de Louvain (PUL) En ligne (consulté le 22/02/21) :
<https://bit.ly/3h3eRsA>
- › Université de Liège, SMART - Système Méthodologique d'Aide à la Réalisation de Tests (2019). *Guide méthodologique. Création d'une évaluation de qualité*. En ligne (consulté le 08/02/2021):
<https://bit.ly/3vHTWj4>
- › Université Libre de Bruxelles, Les fiches-outils du Centre d'appui pédagogique (N.D). La « triple concordance » dans une action de formation. En ligne (consulté le 15/02/2021):
<https://bit.ly/3eU95aa>
- › University of Waterloo, Designing Multiple-Choice Questions, Center for Teaching Excellence. En ligne (consulté le 08/03/2021):
<https://bit.ly/3xIW3Fb>
- › UYTTEBROUCK, E. (2004a). Seuil de réussite à 10 ou à 12 ? Le cas particulier des QCM. Les fiches-outils du Centre d'appui pédagogique, Université Libre de Bruxelles. En ligne (consulté le 15/02/2021):
<https://bit.ly/2RsQMRq>
- › UYTTEBROUCK, E. (2004b). QCM et points négatifs. Les fiches-outils du Centre d'appui pédagogique, Université Libre de Bruxelles. En ligne (consulté le 15/02/2021):
<https://bit.ly/3uhoFD8>
- › UYTTEBROUCK, E. (2020). Présentation QCM et correction des effets du hasard. Présentation au Conseil des Études. Centre d'appui pédagogique, Université Libre de Bruxelles.
- › WILLIAMS J. (2006). Assertion-reason multiple-choice testing as a tool for deep learning: a qualitative analysis, *Assessment & Evaluation in Higher Education*, 31 (3), pp. 287-301. En ligne (consulté le 15/02/2021):
<https://bit.ly/3gZwvte>

- › ZIMMARO, D. (2016). Writing good multiple-choice exams, The University of Texas at Austin, Faculty Innovation Center. En ligne (consulté le 15/02/2021):

 <https://bit.ly/3nOmLYc>

