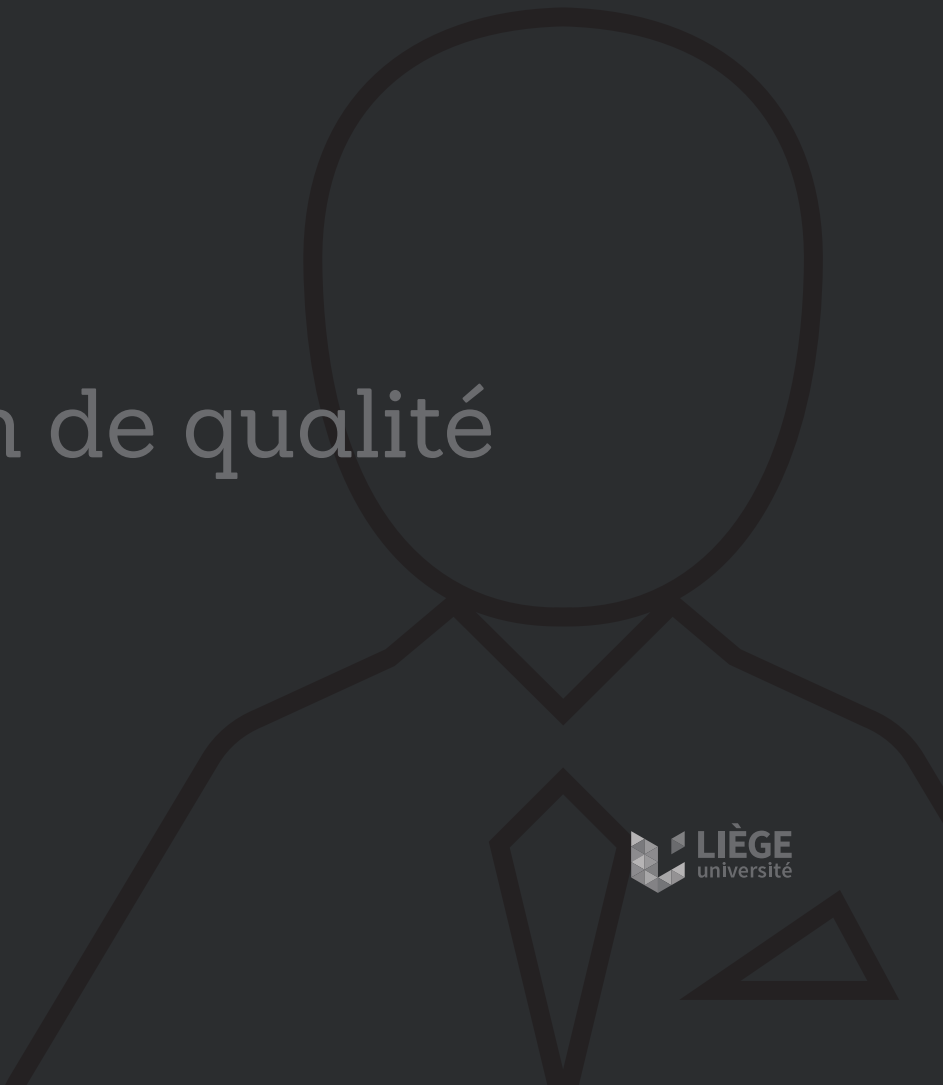
SMART

Système Méthodologique
d'Aide à la Réalisation de Tests

Guide méthodologique

Création d'une évaluation de qualité



LIÈGE université

SMART

Système Méthodologique
d'Aide à la Réalisation de Tests

Guide méthodologique

Création d'une évaluation de qualité

IFRES LIÈGE université

Une unité de soutien de l'IFRES • Université de Liège


QUESTIONNAIRE À CHOIX MULTIPLE AVEC DEGRÉS DE CERTITUDE
QUESTIONNAIRE À CHOIX MULTIPLE
Consignes de marquage
En cas d'erreur de marquage,
ne raturez pas
mais utilisez la seconde ligne
pour cocher la réponse définitive
bic noir ou bleu
Ne faites pas
1
2
3
4
5
6
7
8
9
0
C
SEND
LINK

Guide méthodologique

Création d'une évaluation de qualité

Depuis de nombreuses années, le Système Méthodologique d'Aide à la Réalisation de Tests de l'Université de Liège intervient dans la correction automatisée d'évaluations (formatives ou certificatives) à l'aide d'outils technologiques et de logiciels spécifiquement dédiés. Cependant, en amont de la correction de vos épreuves, une série de règles et de procédures sont à suivre afin de réaliser une évaluation de qualité.

Ce guide méthodologique a pour vocation de vous en présenter les principales et de vous fournir les outils nécessaires à la réalisation d'un examen composé soit de questions à choix multiples, y compris les vrai-faux et les questions à choix large, soit de questions à réponses ouvertes courtes ou moyennes.

Ainsi, de nombreux soucis en termes de validité et de fidélité de votre épreuve pouvant survenir *a posteriori* pourront être évités.

Ce livre est composé de deux grandes parties. La première se centre sur :

- l'importance de la validité d'une évaluation;
- la méthodologie pour créer une évaluation de qualité avec la table de spécification comme élément central;
- les différents types d'évaluations non-exhaustifs et un synoptique des caractéristiques docimologiques des modalités de questionnement.

La deuxième partie est consacrée plus spécifiquement aux évaluations standardisées. Nous y aborderons :

- la définition d'une question à choix multiple (QCM);
- les règles de rédaction d'une QCM;
- les caractéristiques optionnelles (solutions générales implicites, degrés de certitude, sévérités);
- les barèmes de correction;
- le processus de correction d'une épreuve via le SMART;
- les résultats fournis par le SMART;
- les feedbacks SMART à destination des étudiants.

Bonne lecture !

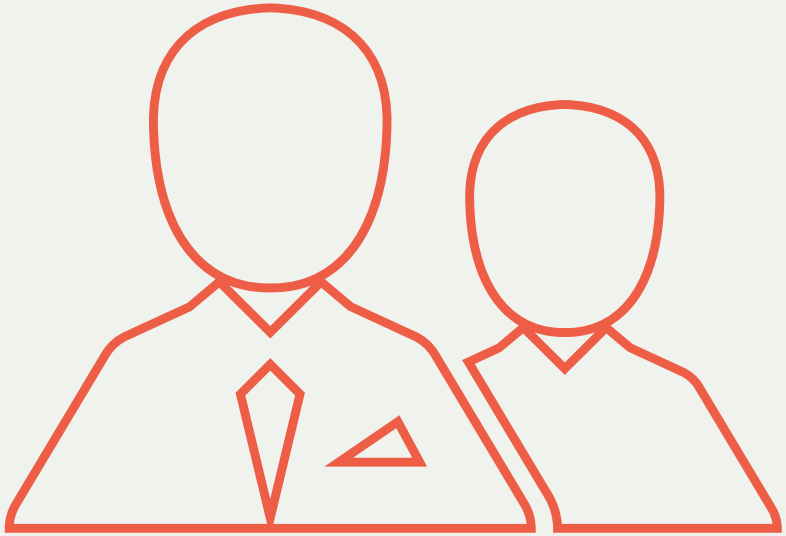


Table des matières

L'évaluation : le défi de la qualité 6
Le cycle qualité du SMART dans la réalisation de tests 11
La table de spécification 12
- Catégories de performance 13
- Modalités de questionnement 18

Synoptique des caractéristiques docimologiques des modalités de questionnement 25
L'approche par compétences 27
Guide de création et de correction d'épreuves QCM 31
- Qu'est-ce qu'une QCM ? 31
- Les solutions générales implicites (SGI) 32
- Les degrés de certitude (DC) 33
- Pourquoi nos procédures qualité ne prévoient-elles qu'une seule réponse correcte par question ? 34
- Rédaction des questions 35

Information et formation des étudiants 40
Le test en pratique 42
- La commande des formuLOMs 42
- Les modalités d'utilisation des différents types de formuLOMs 43
- La passation 47
- Documents à remettre au SMART et informations à communiquer lors du dépôt des formuLOMs 48

La correction 48
- Barèmes de correction 49
- Utilisation de niveaux de sévérités différents plutôt qu'un ajout de points 51
- Correction automatisée et contrôles qualité 54
- Envoi des résultats et validation de l'épreuve 58

Exportation des notes via myULiège (uniquement pour les enseignants de l'Université de Liège) 59
Les feedbacks aux étudiants 59
Les feedbacks aux professeurs 60

L'évaluation : le défi de la qualité

Selon Nitko et Brookhart (2011, p.3), une évaluation se définit comme étant « *un processus pour obtenir de l'information qui sera utilisée pour prendre des décisions à propos d'étudiants, de programmes de cours et d'écoles, ainsi qu'à propos de politiques éducatives* ».

Dès lors, que l'évaluation soit à visée formative ou certificative, elle se doit d'être la plus valide et la plus fidèle possible afin que les mesures recueillies reflètent au mieux les acquis d'apprentissage (dans le cas des étudiants) que l'on souhaite appréhender lors du test que l'on administre aux étudiants. L'idée théorique est de faire en sorte que le différentiel entre l'apprentissage vrai (réel de l'étudiant) et l'apprentissage mesuré par le dispositif d'évaluation soit le plus faible possible.

Cependant, créer une épreuve de qualité afin d'appréhender tous les concepts visés et atteindre tous les objectifs fixés n'est pas chose aisée si l'on ne tient pas compte d'un certain nombre de caractéristiques garantes d'une évaluation de qualité. Celles-ci sont définies à travers le concept de validité.

En 1999, ce concept a été redéfini par un ensemble de structures majeures. Il s'agit de l'*American Educational Research Association*, de l'*American Psychological Association* et du *National Council on Measurement in Education* qui lui donneront la définition suivante : l'analyse de la validité est un processus dont le but est d'accumuler des preuves démontrant le caractère approprié des inférences qui sont faites à partir des résultats d'un test donné. La validité a trait, dès lors, au degré de conviction que l'on peut avoir – suite à l'analyse de ces preuves – quant à l'exactitude des interprétations qui sont faites à partir du test et quant au caractère approprié de la façon dont celles-ci sont réalisées.

Un des premiers éléments marquant de cette définition est que la validité ne s'applique pas à un test décontextualisé mais bien en usage et que l'interprétation des résultats d'un test se fait en contexte. En conséquence :

- C'est bien l'utilisation d'un test et l'interprétation des résultats qui en sont issus, qui sont les objets de la validation, et non le test en lui-même. Dans ce cadre, malgré un usage courant, dire qu'un test est valide sans en préciser le contexte de validation n'a pas de sens.

- Si l'utilisation et l'interprétation sont les objets de l'étude de validité, le type et la quantité de preuves nécessaires varient en fonction du contexte. Les critères applicables à un test formatif ne sont pas nécessairement identiques à ceux que l'on applique lors d'un test certificatif.
- Si le concept de validité est un concept qui apparaît comme étant relativement simple, l'étude de la validité est cependant quelque chose de nécessairement complexe qui prend en compte l'écologie du milieu dans lequel le test a été mis en œuvre.

Cette définition s'applique et reste tout à fait cohérente quel que soit le type d'évaluation concerné, que ce soit dans le cadre de la théorie classique des tests ou dans le cadre de l'évaluation dans une approche par compétences.

Traditionnellement, il sera nécessaire de combiner plusieurs éléments d'analyse pour mener une étude de validité. Selon Sireci (2009), six types de preuves doivent être recueillies : celles portant sur le contenu, le process, le critère, la fidélité, la conséquence et la praticabilité.

La **preuve axée sur le contenu** réfère à la manière dont les réponses d'un étudiant à une évaluation donnée reflètent réellement les connaissances que celui-ci a du contenu de la matière abordée dans le test. En d'autres termes, il s'agit du différentiel entre une connaissance mesurée dans un test — pour un individu donné — et la connaissance réelle que cet individu a de la matière. Cette connaissance réelle ne peut être mesurée directement : il s'agit d'une abstraction. Dans ce contexte, les docimologues vont s'intéresser à la manière dont le test rend compte de la matière qui doit être évaluée. Traditionnellement, dans le monde de l'éducation, cette étude porte sur l'alignement entre les éléments du curriculum (ou du référentiel de compétences/de formation), les éléments enseignés lors de la formation et les éléments réellement évalués lors de l'examen.

La **preuve axée sur le process** est définie comme étant le recueil de preuves concernant l'alignement entre les processus mentaux qui ont fait l'objet d'un enseignement, ceux visés par l'évaluation et ceux réellement mis en œuvre par l'étudiant.

La **preuve basée sur le critère externe** s'intéresse à la relation entre le score à un test donné et d'autres variables identifiées comme étant des mesures du même sujet que celui abordé par le test. Par exemple, pour vérifier la validité d'un dispositif d'évaluation de l'enseignement par les étudiants, on corrélera cette mesure avec d'autres indicateurs de la qualité de l'enseignement,

comme par exemple, l'avis de pairs, l'analyse du portfolio de l'enseignant…

L'**analyse de la fidélité d'un test** peut se faire de quatre manières différentes. Il s'agit de :

- La fidélité test-retest : pour évaluer sa fidélité, le test est administré deux fois, à deux moments différents et doit produire le même score pour être valide. La fidélité test-retest est cependant difficile à mettre en œuvre dans un contexte éducatif. En effet, on constate chez les étudiants un effet d'apprentissage en lien avec la première passation du test qui influence les résultats lors de la deuxième passation.
- La fidélité entre formes parallèles : on compare deux tests créés à partir du même contenu (exemple : en séparant aléatoirement les items d'un même test) et passés en même temps par les mêmes étudiants. Les deux formes doivent donner les mêmes résultats pour que le test soit valide. Ce type de fidélité nécessite un certain nombre d'items tirés de manière aléatoire. C'est possible dans le cas de QCM à condition de mettre en place des modèles statistiques appelés « *modèles de réponse à l'item* » pour s'assurer que les formes parallèles soient d'une difficulté similaire. Toutefois, ce type de validité est malaisé à évaluer dans le cadre d'épreuves axées sur une approche par compétence, nécessitant l'utilisation de situations complexes. On pourra toutefois éventuellement s'appuyer sur des situations faisant partie de la même famille. L'analyse ne serait donc pas faite entre des formes parallèles de tests, mais bien entre des formes parallèles de situations issues d'une même famille.
- La fidélité par cohérence interne : cette forme de fidélité est utilisée pour juger de la consistance des résultats à travers l'analyse des items d'un même test, par exemple à l'aide de l'analyse du *r.bis* ou de l'*alpha de Cronbach*.
- La fidélité inter-juges : ce type de fidélité est mesuré lorsque deux correcteurs indépendants évaluent la même production. Différents indicateurs statistiques, comme par exemple le *coefficient de Kappa*, peuvent être appliqués et fournir une information précise de la fidélité inter-juges.

La **preuve basée sur la conséquence** évalue les effets souhaités et non souhaités de l'application d'un test ou d'un dispositif d'évaluation. Le type de méthodologie permettant de récolter ce type de preuve va de la mesure d'impact de l'évaluation *a posteriori* des inférences qui ont pu être faites sur base des résultats de l'évaluation, mais également à l'observation du comportement des individus (insatisfaction, déprime…) lors de l'évaluation et après celle-ci.

La **preuve axée sur la praticabilité** s'intéresse à l'efficience des outils d'évaluation. Sont-ils réplicables, en des coûts raisonnables, à d'autres publics ou dans d'autres contextes ?

Ces différentes sources de preuves offrent un cadre riche pour instruire sur la qualité d'un test et de son utilisation ainsi que pour récolter divers éléments permettant d'étudier cette délicate question : le niveau de validité. Ce cadre de travail nous pousse à adopter différentes stratégies pour analyser la validité d'un test appliqué dans un certain contexte. Toutefois, ces recommandations n'ont pas un caractère clairement prescriptif. Le manque de règles fortes et prescriptives peut sembler insatisfaisant auprès du néophyte. Il est vrai que des questions cruciales restent (un peu) en suspens. Par exemple, « *Quand peut-on considérer que nous possédons suffisamment de preuves pour acter la validité ?* », ou encore, « *Que faire lorsqu'un des types de preuves est insatisfaisant mais que tous les autres plaident pour une excellente validité ?* ». Il nous faut toutefois reconnaître que les sciences de l'éducation et l'édumétrie sont trop complexes pour proposer des tests statistiques simples inférant la validité ou encore pour fournir un livre de recettes pour conduire des études de validité.

Étudier la validité n'est… *simplement pas simple*. Et toute idée simpliste à ce sujet mènerait à des caricatures non souhaitables.

Pour atteindre ou améliorer la validité d'un dispositif d'évaluation, quelques règles issues de la docimologie ont été développées dans le but d'aider tous les formateurs et personnes devant réaliser des évaluations.

C'est dans ce cadre que, depuis 2003, le SMART a développé un cycle de « **Construction et Gestion Qualité de Tests Standardisés** » composé de **huit étapes** servant de guide pour l'élaboration d'une évaluation à l'aide de questions à choix multiple et pour la rédaction de ces questions.

Ce cycle qualité, conçu à l'origine pour la réalisation d'évaluations standardisées utilisant des questionnaires à choix multiple (QCM), peut cependant être adapté pour la mise sur pied d'autres modalités de questionnement tout en restant toujours reconnu pour ses qualités docimologiques.

Ce sont donc ces huit étapes que nous vous proposons d'appliquer et de décrire dans la suite du présent guide méthodologique.

Le cycle qualité du SMART dans la réalisation

1. Analyse

Lors de cette étape, nous garantissons la validité de l'épreuve que l'on souhaite développer grâce à des procédures qui vont nous permettre de relever les points de matière que l'on souhaite évaluer, ainsi que les catégories de performance (souvent, selon la taxonomie de *Bloom* : connaissance, compréhension, application, synthèse…) que l'on souhaite voir acquises par les étudiants.

2. Design

Il s'agit ici de choisir les modalités de questionnement que l'on souhaite utiliser. Celles-ci sont définies à partir des points évalués et des catégories de performance définis à l'étape d'analyse.

3. Questions

C'est à cette étape que la rédaction des questions de l'épreuve va prendre place. Après avoir défini la matière à évaluer ainsi que la façon dont on veut l'évaluer, le processus de construction des questions va pouvoir débuter. Différents contrôles de la qualité *a priori* des items rédigés pourront être réalisés grâce, notamment, à l'utilisation des règles de rédaction ainsi que via la validation par des experts.

4. Training

Cette étape permet au responsable d'évaluation de proposer des tests d'entraînement aux étudiants, leur permettant ainsi de se familiariser avec la procédure d'évaluation, mais également de déjà parcourir la matière et de l'étudier une première fois.

5. Passation

C'est à ce niveau que va prendre place l'évaluation à proprement parler, que l'on va associer différentes questions à un test particulier et que l'on va soumettre le questionnaire aux étudiants. Lors du choix des questions, il sera important de s'assurer que l'ensemble de celles-ci couvre toute la matière ainsi que les catégories de performance définies lors de l'étape d'analyse.

de tests

6. Correction

Suite à la passation de l'épreuve, une étape de correction est nécessaire pour pouvoir attribuer une note à l'étudiant. Lors de cette étape, il existe également une série de contrôles qualité des questions, *a posteriori* cette fois, d'indices statistiques et classiques de cohérence interne, permettant de vérifier la qualité de l'épreuve passée par les étudiants. Grâce à ces différents indices, des actions correctives pourront ensuite être prises dans le souci d'améliorer le niveau de qualité de l'évaluation.

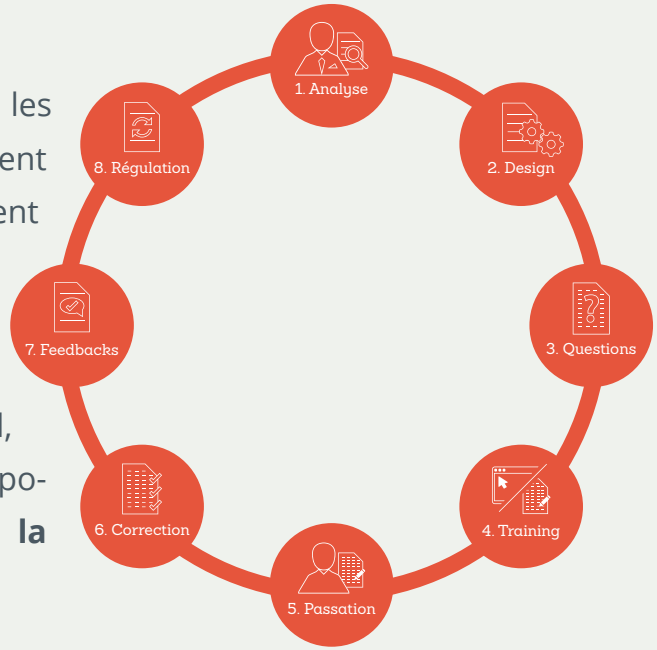
7. Feedbacks

Une fois l'épreuve validée et entérinée, consécutivement aux différentes modifications qui ont pu être faites, la mise à disposition de *feedbacks*, de rétro-informations aux étudiants leur permet d'avoir un retour sur l'épreuve qu'ils ont passée et ainsi, d'avoir un aperçu de leurs points forts et de leurs faiblesses.

8. Régulation

Cette étape permet de récolter de l'information sur le processus d'évaluation mis en œuvre et de procéder à des réajustements éventuels après avoir obtenu différents avis de la part des personnes ayant passé l'épreuve. Cette étape permet donc d'améliorer encore un peu plus le processus de construction du test.

Parmi ces huit étapes, les deux premières permettent de construire un élément central à développer pour garantir les qualités docimologiques des évaluations mises sur pied, indépendamment du dispositif d'évaluation choisi : **la table de spécification**.



Le SMART propose donc de développer la façon de construire cette table de spécification, de manière à fournir une explication commune à toute réalisation de processus évaluatif.

La table de spécification

Un élément central dans la construction d'évaluations de qualité

« Pour concevoir une tâche ou une situation-problème permettant d'inférer une compétence, il faut « interpréter *» l'énoncé de cette compétence en termes opérationnels : cela exige de traduire les composantes ou les capacités qui sous-tendent cette compétence en ressources à mobiliser. C'est là un enjeu fondamental de l'évaluation »* (p. 120, Scallon, 2004).

L'utilisation d'une table de spécification nous garantit la validité de l'épreuve que l'on souhaite développer grâce à des procédures qui vont nous permettre de relever les différents points cruciaux en vue de l'élaboration d'une évaluation de qualité :

- Les points de matière vus lors du cours et à évaluer ;
- La priorité avec laquelle ces points de matière seront évalués ;
- Les catégories de performance que l'on souhaite voir acquises par les étudiants, en lien avec les points de matière ;
- La modalité de questionnement la plus adaptée aux objectifs d'évaluation.

De la sorte, des balises seront posées afin de pouvoir ensuite débuter la construction de notre évaluation de manière optimale et d'en renforcer les validités de contenu et de process. Nous saurons que l'analyse de la matière à évaluer, ainsi que la mise en place du design de l'évaluation, permettront de créer une épreuve valide et rencontrant des qualités docimologiques indéniables.

La première chose à faire lorsque l'on veut créer une table de spécification de manière optimale est de **lister les points de matière, les points à évaluer (PE)** qui ont été abordés lors du cours. Il peut s'agir ici de partir de la table des matières, de normes (inter)nationales ayant été fixées par ailleurs, ou encore d'un référentiel de compétences.

Catégories de performance

La seconde sous-étape consiste à **lister les catégories de performance (CP)** qui seront visées par l'évaluation. Ici, on se réfère régulièrement à la *taxonomie de Bloom* pour les savoirs, mais toute autre taxonomie, pour peu qu'elle ait été validée, peut être utilisée.

Pour les savoir-faire et savoir-être, nous proposons l'utilisation des taxonomies de *Jewett* et *Krathwohl*, qui sont les plus connues, mais d'autres taxonomies sont tout aussi utilisables.

La *taxonomie développée par Bloom* (1956) peut être décomposée en 6 catégories de performance, allant d'un niveau très concret, à un niveau plus abstrait :

1. **Connaissance** : Savoir retransmettre ou reproduire avec justesse toute information, connaissance ou procédure préalablement acquise (donc, ce n'est pas le mécanisme de l'acquisition des connaissances, mais le fait de les avoir acquises pour pouvoir les restituer qui est évalué ici). C'est le niveau le plus concret de la taxonomie.
2. **Compréhension** : Être capable de saisir le sens littéral d'une communication, d'exprimer avec ses propres mots ce qui est acquis.
3. **Application** : Utiliser les connaissances acquises antérieurement (dont les règles de procédure) dans de nouvelles situations pour tenter de résoudre des problèmes de la meilleure façon ou de façon univoque.
4. **Analyse** : Morceler ou découper un objet ou de l'information selon ses parties, les examiner (tout en tentant de les comprendre ou d'en comprendre le fonctionnement ou la structure) en isolant les causes, en faisant des inférences, afin de pouvoir généraliser.
5. **Synthèse** : Mettre en application un ensemble de connaissances et d'habiletés afin de créer un objet nouveau, cohérent et original.
6. **Évaluation** : Porter un jugement sur la valeur de quelque chose en se basant sur ses connaissances, ses méthodes et ses valeurs afin de proposer un produit entier et nouveau, selon un but précis et des protocoles établis. C'est le niveau le plus abstrait de la taxonomie.

Taxonomie des apprentissages cognitifs de Bloom (1956)

Connaître	Comprendre	Appliquer	Analyser	Synthétiser	Évaluer
Définir, énumérer, mémoriser, ordonner, rappeler, reconnaître, répéter	Comparer, dire en ses mots propres, discuter, décrire, reformuler, traduire	Calculer, employer, formuler, manipuler, résoudre, traiter, utiliser	Différencier, déduire, examiner, identifier, inférer, organiser, trouver	Adapter, composer, créer, développer, planifier, produire, structurer	Argumenter, choisir, critiquer, défendre, estimer, juger, justifier

Jewett (1974) propose une *taxonomie des habiletés motrices*. Celle-ci se décline en 8 catégories. Nous vous livrons les définitions proposées par Lasnier en 2000 :

Taxonomie du domaine moteur (Jewett, 1974)

Mouvements généraux

Percevoir	Imiter	Exécuter
Identifier, découvrir, reconnaître, discriminer	Reproduire, imiter, mimer	Réaliser, démonter, coordonner un modèle

Mouvements ordonnés

Adapter	Raffiner
Ajuster, appliquer, employer, utiliser	Perfectionner, contrôler, synchroniser, améliorer, systématiser, réaliser d'une manière facile et efficace

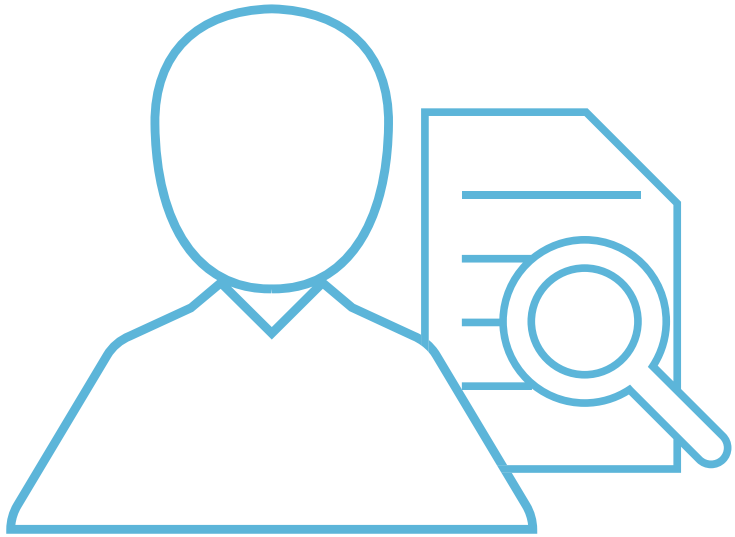
Mouvements créatifs

Varier	Improviser	Composer
Modifier, changer, diversifier	Interpréter, improviser, anticiper	Symboliser, composer

1. **Percevoir** : Reconnaître des mouvements, des positions, des modèles et habiletés par les organes sensoriels.
2. **Imiter** : Duplication d'un modèle moteur ou d'une habileté comme étant le résultat de la perception.
3. **Réaliser un modèle (Exécuter)** : Arrangement et utilisation de diverses parties du corps d'une manière harmonieuse pour réaliser un mouvement ou une habileté.
4. **Adapter** : modification d'un mouvement ou habileté modèle afin de rencontrer certaines demandes spécifiques de tâche.
5. **Raffiner, perfectionner** : Acquisition de contrôle facile et efficiente dans la réalisation d'un mouvement ou d'une habileté modèle sous l'action d'un processus de perfectionnement par : élimination des mouvements parasitaires ; maîtrise des relations spatiales et temporelles ; réalisation habituelle dans des conditions plus complexes.
6. **Varier** : Invention ou construction d'options nouvelles dans la réalisation de mouvements ou d'habiletés.
7. **Improviser** : Création ou initiation de mouvements originaux ou de combinaisons de mouvements.
8. **Composer** : Création de mouvements uniques en fonction d'une intention.

La *taxonomie de Krathwohl* (1964) quant à elle, a été proposée en collaboration avec Bloom et est liée au domaine affectif. Celle-ci comprend 5 niveaux décomposés en 2 ou 3 sous-niveaux :

1. **Réception (présence)** : Sensibiliser l'individu à l'existence de certains phénomènes et certains stimuli, c'est-à-dire l'inciter à les recevoir ou à y faire attention.
2. **Réponse** : Réponses qui suivent la simple attention prêtée aux phénomènes. On souhaite qu'un individu soit suffisamment engagé dans un sujet, un phénomène ou une activité pour chercher à le découvrir et avoir plaisir à l'approfondir.



3. **Valorisation** : Comportement qui est assez solide et stable pour prendre les caractéristiques d'une croyance ou d'une attitude. L'individu manifeste ce comportement avec suffisamment de cohérence, dans les circonstances appropriées, pour que l'on estime qu'il détient une valeur. Intériorisation d'un ensemble de valeurs spécifiques idéales : le comportement est motivé, non par le désir de plaire ou d'obéir, mais par l'engagement individuel à la valeur fondamentale déterminant le comportement.
4. **Organisation** : Organiser les valeurs en système, déterminer les inter-relations qui existent entre elles, établir celles qui sont dominantes et plus profondes.
5. **Caractérisation par une valeur ou un système de valeurs** : Les valeurs ont une place dans la hiérarchie des valeurs de l'individu. Elles sont organisées en une sorte de système intrinsèquement cohérent. Elles ont réglé le comportement de l'individu assez longtemps pour que celui-ci s'y soit adapté.

Taxonomie affective de Krathwohl (1964)

Réception			Réponse		
Conscience	Volonté de recevoir	Attention dirigée	Assentiment	Volonté de répondre	Satisfaction de résoudre
[Différencier, séparer, isoler] * [des vues, des sons, des évènements]	[Accumuler, choisir, combler, accepter] * [des modèles, des exemples, des configurations]	[Choisir, répondre corporellement écouter, contrôler] * [des alternatives, des réponses, des nuances]	[Se conformer, suivre, confier, approuver] * [des directions, des instruments, des lois, des démonstrations]	[Offrir spontanément, discuter, pratiquer, jouer] * [des instruments, des jeux, des œuvres, des parodies]	[Applaudir, acclamer, passer ses loisirs à] * [des discours, des pièces, des présentations]

Valorisation			Organisation		Caractérisation (valeurs)	
Acceptation d'une valeur	Préférence pour une valeur	Engagement	Conceptualisation	Organisation d'un système de valeur	Disposition généralisée	Caractérisation
[Améliorer sa compétence en, augmenter des quantités de, renoncer, spécifier] * [membres d'un groupe, productions artistiques, productions musicales, amitiés personnelles	[Assister, aider, encourager] * [des artistes, des projets, des points de vue, des arguments]	[Nier, protester, débattre, argumenter] * [des déceptions, des inconséquences, des abdications, des irrationalités]	[Abstraire, comparer, discuter, théoriser] * [des buts, des codes, des standards, des paramètres]	[Harmoniser, organiser, définir] * [des systèmes, des approches, des critères, des limites]	[Réviser, changer, compléter, réclamer] * [des plans, des comportements, des méthodes, des efforts]	[Être bien évalué par ses pairs, être bien évalué par ses supérieurs] * [pour humanitarisme, pour morale, pour intégrité]

Comme nous venons de l'expliquer, à ce stade, la table de spécification peut être considérée comme un tableau à double entrée comprenant d'une part les points de matière qui ont été abordés par la formation et qui pourraient être évalués et d'autre part, des catégories de performance en lien avec diverses taxonomies. Il nous faut à présent en **déduire les objectifs d'apprentissage en croisant**, lorsque c'est pertinent, **ces deux niveaux d'analyse**. Cette étape est primordiale pour la première partie de la réalisation de la table de spécification.

Ce travail étant effectué, il ne reste plus qu'à **attribuer des priorités à chaque objectif d'apprentissage**, en fonction de l'importance que l'on accorde à chacun d'eux.

Formation	CP 1	CP 2	CP 3
Chapitre 1			
Section 1.1			
Contenu 1.1.1 : **PE 1**	×	×	
Contenu 1.1.2 : **PE 2**		×	
Section 1.2	×		
Contenu 1.2.1 : **PE 3**	×		×
Contenu 1.2.2 : **PE 4**		×	
Chapitre 2			

Objectif d'apprentissage

Modalités de questionnement

Ensuite, il est nécessaire de donner une troisième dimension à ce travail d'analyse. Celle-ci est issue d'une réflexion sur le type de **modalité de questionnement particulière (MQ)** que l'on va pouvoir utiliser pour évaluer les divers objectifs d'apprentissage.

Pour ce faire, il faut nécessairement tenir compte d'une part des :

- Objectifs d'apprentissage visés (et uniquement ceux-ci) ;
- Importances relatives des objectifs d'apprentissage visés ;
- Décisions qui doivent être nourries par le processus d'évaluation (principes de faisabilité et de pragmatisme).

D'autre part, il faudra mettre ces divers éléments en lien avec les avantages et inconvénients des différentes modalités de questionnement.

Voici une liste non exhaustive de certaines de ces modalités de questionnement et leurs possibilités d'application, en fonction des objectifs d'apprentissage visés :

1. **Question à choix large (QCL)** : rappel aidé.
2. **Question à choix multiple (QCM)** : reconnaissance, compréhension (si à livre ouvert ou avec des *SGI*), application, analyse (surtout si *SGI* « *manque* » et « *absurdité* »).
3. **Question à réponse ouverte courte (QROC)** : rappel-évocation, compréhension, application, analyse.
4. **Question à réponse ouverte longue (QROL)** : rappel-évocation, compréhension, application, analyse, synthèse (via résolution de problèmes).
5. **Projet** : connaissance, compréhension, application, analyse, synthèse, créativité, évaluation.
6. **Portfolio** : connaissance, compréhension, application, analyse, synthèse, créativité, évaluation.
7. **Simulation** : connaissance, compréhension, application, analyse, synthèse, créativité, évaluation.
8. **Examen clinique objectif structuré (ECOS)** : connaissance, compréhension, application, analyse, synthèse, créativité, évaluation.

Nous vous proposons maintenant de définir les modalités de questionnement les plus classiques et d'en énoncer quelques avantages et inconvénients. Ensuite, nous les classerons dans un tableau à double entrée permettant d'en saisir les qualités et défauts de manière synoptique.

Les questions à choix multiple

Elles consistent en une phrase d'introduction (**l'amorce**), suivie par une liste de solutions proposées parmi lesquelles on trouve **des distracteurs** et **la réponse correcte**. Elles peuvent être affinées par l'utilisation des solutions générales implicites et des degrés de certitude. Elles présentent les qualités suivantes :

- Elles permettent une notation automatique (la rapidité de correction et la fidélité intra-correcteur en sont fortement améliorées).
- Elles permettent d'évaluer plus d'objectifs d'apprentissage que n'importe quel autre type de questions fermées (*vrai-faux* par exemple).
- Elles minimisent l'opportunité de *bluffer* ou d'habiller ses réponses et se centrent sur la lecture et la réponse, pas sur la compétence écrite.
- Elles réduisent les chances de donner la réponse correcte en la devinant (par rapport au *vrai-faux*).
- Elles peuvent donner une indication sur les difficultés rencontrées par un étudiant.
- Elles permettent de poser beaucoup de questions en un temps réduit, ce qui permet d'atteindre une bonne validité de contenu.

Toutefois, elles présentent également certains défauts :

- Elles ne sont pas appropriées pour les catégories de performances les plus élevées, ni pour les savoir-faire et savoir-être.
- Elles ne favorisent ni l'émergence de créativité, ni l'expression de pensée personnelle.
- Mal formulées, elles peuvent être superficielles, triviales et limitées aux connaissances factuelles.
- Elles peuvent pénaliser les meilleurs étudiants, qui ont un esprit trop critique.
- Leur usage exclusif peut entraîner l'apprentissage dans une direction non souhaitable (étude par cœur, manque de compréhension et de liens entre les éléments de la matière…).

Si leur usage exclusif pose des problèmes, on les choisira toutefois de manière privilégiée lorsqu'il s'agira **d'évaluer la connaissance, la compréhension ou l'application**. En effet, les avantages liés à leur correction et à leur capacité à évaluer une matière large en un temps raisonnable en fait des alliées de poids pour maîtriser la validité de contenu et les aspects liés à la fidélité.

Les productions longues

Il s'agit de demander aux étudiants d'écrire une production longue où ils sont libres d'exprimer et d'organiser leurs idées ainsi que l'inter-relation entre celles-ci. Le recours à ces méthodes permet d'évaluer des niveaux taxonomiques plus élevés que les questions à choix multiple mais également de combiner ces derniers. De plus, elles donnent la possibilité d'évaluer les compétences liées à la lecture et favorisent l'expression de la créativité ou, à tout le moins, l'expression personnelle. Elles influencent également les apprentissages puisque les liens entre divers éléments de matière peuvent être faits. Elles restent également peu coûteuses en terme de passation puisqu'elles peuvent se présenter dans le cadre d'un examen papier-crayon classique. On notera toutefois trois défauts majeurs :

- Elles nécessitent un temps de réponse relativement long de la part de l'étudiant. En conséquence, elles sont en nombre limité (seulement quelques questions par évaluation) et présentent souvent un faible niveau en terme de validité de contenu.
- Elles consomment énormément de temps lors de la correction.
- Elles présentent assez souvent une faible fidélité inter- ou intra-correcteur.

Tout comme pour les questions à choix multiple, un usage exclusif peut poser des problèmes difficilement surmontables. On les utilisera de manière privilégiée pour **évaluer des niveaux taxonomiques comme la synthèse, l'évaluation ou la créativité**.

Les résolutions de problèmes

Il s'agit d'utiliser des connaissances et habiletés dans un contexte nouveau où le chemin à suivre, la solution ne peuvent être déterminés de manière automatique par l'étudiant. Les problèmes peuvent être bien ou mal définis, ne comporter qu'une ou plusieurs réponses correctes. On observera essentiellement la maîtrise de la démarche de résolution de problème (par exemple, la démarche *IDEAR* pour *Identifier, Définir, Explorer, Agir, Rétroagir*), même si les résultats peuvent, eux aussi, être évalués. L'avantage de ce type de méthode est qu'il présente **un bon niveau d'authenticité** puisque le problème choisi peut être conçu à partir d'une situation réelle en lien avec le contexte de travail.

L'évaluation des performances

Dans ce contexte, l'étudiant crée un produit ou démontre l'acquisition d'un processus (ou les deux). Les tâches demandées peuvent aller de courtes réalisations (simulations, ECOS) jusqu'à des travaux intégratifs de type portfolio.

Les avantages de cette modalité sont sa capacité à évaluer des objectifs d'apprentissage de très haut niveau (et leur intégration), ainsi que la *capacité à agir*. Elles permettent dès lors d'évaluer des capacités acquises et des savoir-faire comportementaux. C'est incontestablement l'une de leur force. En termes d'évaluation on notera l'implication de l'utilisation de grilles descriptives. Le processus et le résultat de la performance peuvent être mesurés. Cette modalité comporte néanmoins un certain nombre de désavantages. Comme par exemple :

- La difficulté et le temps nécessaire pour la création et la correction de ces types d'épreuves.
- Le temps nécessaire pour la passation de telles épreuves, qui empêche leur multiplication et peut poser des problèmes en termes de validité de contenu.

Tout comme pour les autres modalités d'évaluation, leur utilisation exclusive pose problème. Elles sont toutefois incontournables quand il s'agit d'**évaluer la personne dans sa capacité à agir**. Notre conseil sera donc de **multiplier les différents types de modalités de questionnement** afin de garantir une bonne validité de contenu et d'évaluer des niveaux de performance élevés.

Une fois déterminées, les modalités de questionnement liées avec les objectifs d'apprentissage se présenteront alors de la sorte dans la **table de spécification** composée de différents **trinômes** :

Point à Évaluer (PE)

×

Catégorie de Performance (CP)

×

Modalité de Questionnement (MQ)

Formation : Biologie	Connaissance		Compréhension		Application
	QCM	QROC	QCM	QROL	QCM
I. Composants d'une cellule	2	2	1	2	2
A. Nucleus	1				
B. Cytoplasme	1	1			
C. Membrane cellulaire		1	1		2
II. Cellules animales vs. végétales	2	2	3		
A. Similarités	1	2			
B. Différences					
1. Membrane cellulaire	1	1	3		
2. Production de nourriture					
III. Membrane cellulaire	1	2	2	1	1
A. Diffusion		1	1		
B. Substances diffusées par les cellules	1	2	1		1
IV. Division cellulaire	4		1	1	
A. Phases de division	2				

Ainsi, les objectifs d'apprentissage auront été conçus et mis en relation avec des modalités de questionnement à privilégier. Tout ce travail préliminaire permet donc déjà, *a priori*, d'augmenter les qualités docimologiques de l'épreuve que l'on souhaite réaliser, en tout cas en termes de validité de contenu et de process.

La table de spécification ainsi obtenue sera ensuite utilisée pour piloter le plan d'évaluation, pour communiquer vers les étudiants et permettre également de leur fournir des feedback performants. On voit donc bien l'intérêt primordial que présente la réalisation d'une telle table pour construire des évaluations de qualité.


y = y0 + V0y t + 1/2 at²

Synoptique des caractéristiques docimologiqu

		Praticabilité			Validité de contenu
		Temps de rédaction	Temps de passation	Temps de correction	
Tests standardisés	Vrai-Faux	✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓
	QCM	✓	✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓
	QROC	✓ ✓	✓ ✓ ✓	✓	✓ ✓
Productions longues	QROL	✓ ✓	✗	✗	✓
	Rédaction	✓ ✓	✗ ✗	✗ ✗	✗
	Essai	✓ ✓	✗ ✗	✗ ✗	✗
Performances	Simulation	✗ ✗	✗ ✗	✗ ✗	✗ ✗
	ECOS	✗ ✗ ✗	✗ ✗	✗ ✗	✓
	Portfolio	✗	✗ ✗ ✗	✗ ✗ ✗	✓ ✓

Légende : ✗ ✗ ✗ médiocre ✗ ✗ mauvais ✗ moyen ✓ bon ✓ ✓ très bon ✓ ✓ ✓ excellent

ues des modalités de questionnement

Validité de construct					Fidélité
Connaissance, application	Compréhension, analyse	Synthèse, évaluation	Savoir faire	Savoir être	
✓ ✓ ✓	✓	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✗	✓ ✓
✓ ✓ ✓	✓ ✓	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✗	✓ ✓ ✓
✓ ✓ ✓	✓ ✓ ✓	✗ ✗ ✗	✗ ✗ ✗	✗ ✗ ✗	✓
✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓	✗ ✗ ✗	✗
✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓	✗ ✗ ✗	✗
✓ ✓ ✓	✓ ✓ ✓	✓ ✓ ✓	✓	✗ ✗ ✗	✗
✓	✓	✓	✓ ✓ ✓	✓ ✓ ✓	✗
✓	✓	✓	✓ ✓ ✓	✓ ✓ ✓	✗
✓	✓	✓	✓ ✓ ✓	✓ ✓ ✓	✗


21-1
2002
21-2
2002
S.1-2
22-1
2003
22-2
2003
SPS.
23-1
2004
23-2
2004
SP.

L'approche par compétences

Grâce à la table de spécification créée, nous pouvons avoir une vue d'ensemble des **points de matière à évaluer** ainsi que des **catégories de performance visées**.

En fonction des objectifs poursuivis, il sera possible d'opter pour deux grands types de dispositifs évaluatifs : une **évaluation standardisée** ou une **évaluation de performance**.

Jusqu'à il y a une quinzaine d'années, l'habitude était de réaliser des **évaluations standardisées** appréhendant si les connaissances avaient bien été transmises auprès des étudiants.

Ce type d'évaluation prend place lorsque « *le même test, le même matériel et les mêmes modalités d'administration pour tous les étudiants* » sont utilisés (Nitko et Brookhart, 2011, p.89).

C'est dans cette optique que des évaluations centrées la plupart du temps sur des questionnaires (QCM et/ou vrai-faux) étaient mises sur pied, sans nécessairement se soucier d'évaluer des compétences plus complexes que de la connaissance et/ou de l'application. Il est donc difficile, si pas impossible, en utilisant le testing standardisé, d'évaluer des performances plus complexes, d'évaluer une compétence dans son ensemble, en lien avec d'autres aptitudes nécessaires à réaliser une tâche complexe.

C'est le cas notamment lorsque nous procédons à une évaluation à l'aide de questionnaires à choix multiple via le SMART.

Depuis 15 ans cependant, la notion de compétence est devenue omniprésente dans le monde professionnel et dans le domaine de l'éducation. Cela a eu pour conséquence que l'enseignement, qui avait pour principale fonction de transmettre les connaissances, vit une mutation menant à une approche basée sur la capacité d'agir en situation complexe. Ce changement de paradigme pédagogique induit par l'**approche par compétences** entraîne la modification de la façon d'aider méthodologiquement les acteurs impliqués dans la réalisation de leurs évaluations.

Dans ce contexte, évaluer les apprentissages est une tâche qui devient de plus en plus difficile pour les enseignants. En effet, comme le souligne Scallon (2004), « *la méthodologie de l'évaluation relève d'un esprit qui va bien au-delà du principe consistant à décrire au mieux ce dont un individu est capable* » (p.11).

L'observation des acquis cognitifs des étudiants n'est donc plus suffisante dans une telle perspective. D'autres caractéristiques accompagnent cette observation telles que la motivation, le degré de confiance, l'engagement, la conscience des difficultés qu'il reste à surmonter…

Face à de telles évolutions, la démarche de l'évaluation a donc été repensée. Un courant nord américain décrit ces nouvelles méthodes en axant l'argumentation sur le concept d' « *assessment* », déployé en « *performance assessment* », « *authentic assessment* » et « *alternative assessment* », chacune de ces expressions renvoyant à une caractéristique de la démarche novatrice.

Parcourons brièvement ces quelques concepts afin de mieux comprendre les traits de cette nouvelle démarche :

Le concept d' « *assessment* » : Lind & Grund (2000) en donnent une définition qui inclut le recours à une diversité de méthodes quantitatives et qualitatives, l'aspect novateur résidant selon nous dans l'intégration de ces méthodes en vue d'aboutir à un jugement reposant sur un **faisceau d'informations** permettant d'assurer la qualité de l'évaluation d'une compétence.

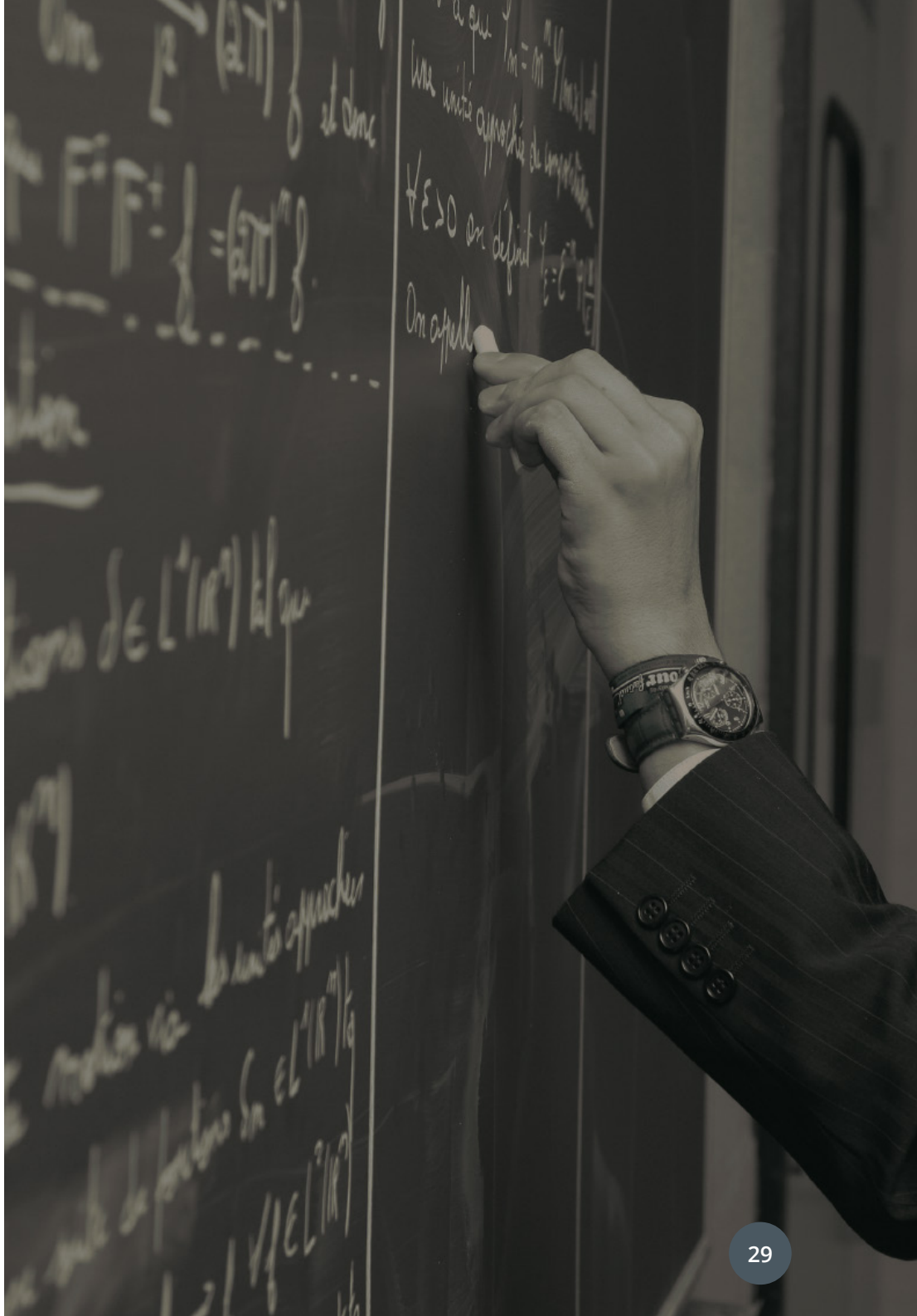
Le concept de « *performance assessment* » : Ces dernières années, on a vu apparaître une tendance visant à mettre les individus en situation de résolution de problèmes leur imposant la construction de réponses élaborées seuls ou en groupes. La complexité des tâches à effectuer peut engendrer un processus évaluatif qui peut s'étaler sur des mois (projets de fin d'études, home exams…) et qui amène l'évaluateur à s'interroger **autant sur les processus que sur le résultat final** (Resnick et Resnick, 1989 ; Portland, 1990 ; Culham et al., 1993 ; Nitko, 1996).

Le concept d' « *authentic assessment* » : En réaction aux reproches adressés aux évaluations objectives, souvent formellement éloignées des réalités de la vie de tous les jours et peu significatives pour les étudiants, on a vu émerger ces dernières années des principes et des procédures docimologiques intégrant des situations **proches de la vie réelle** qui contribuent à donner du sens aux démarches évaluatives (Archbald, 1991 ; Hernandez-Gantes et Phelps, 1996).

Le concept d' « *alternative assessment* » : Les nouveaux défis du renouveau en évaluation imposent le développement de nouvelles méthodes évaluatives (Nitko, 1996) comme les **échelles descriptives** appliquées à l'approche par compétences (De Bal et al., 1976 ; Scallon, 1999, 2004).

Dans cette nouvelle perspective, les outils d'évaluation ont dû s'adapter, se transformer afin de rester en cohérence avec l'approche par compétences. Le questionnaire à choix multiple ne permettant plus de remplir les fonctions demandées par le nouveau champ pédagogique, à savoir évaluer les performances complexes comme les savoirs qui sont mobilisés pour résoudre des problèmes issus de situations de la vie quotidienne et professionnelle, il a donc fallu trouver d'autres instruments pour identifier et développer les compétences des étudiants. Dans ce contexte, le SMART a développé **un processus de qualité pour la correction de questions à productions longues** qui repose sur la dématérialisation des copies, l'accompagnement d'une grille d'évaluation, l'anonymisation des feuilles et la correction par question. Ce processus assure une plus-value en termes d'objectivité, de fiabilité des notes et d'équité qui sont des critères essentiels pour garantir la validité de l'épreuve.

Cependant, il faut garder à l'esprit que ces **deux modes d'évaluation restent complémentaires** aux autres existants et qu'**aucun ne prime sur les autres**. **Les défauts des uns sont compensés par les qualités des autres**. C'est d'ailleurs pour cette raison que nous rappelons (encore) toute l'importance de la création d'une table de spécification afin de pouvoir choisir le type d'évaluation le plus adapté à ce que l'on souhaite évaluer.




39
40

Guide de création et de correction d'épreuves QCM

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 »*.

Ses trois composantes sont :

1. **La consigne** : elle décrit :
   - La question : nombre de réponse(s) correcte(s).
   - Le mode de réponse : une seule ou plusieurs réponses correctes.
   - Les principes de notation : nombre de points attribués/retranchés en cas de réponse correcte, incorrecte ou en cas d'omission.
   - Le barème des conséquences : le poids des questions dans le test est-il identique ou différent ?
2. **L'amorce** : elle définit le problème, elle pose la question.
3. **Les solutions proposées** : elles comprennent la solution correcte et des solutions incorrectes (des *distracteurs*).

Les solutions générales implicites (SGI)

Selon les cas, il est possible d'utiliser les solutions générales implicites (SGI) en complément des solutions proposées dans une question à choix multiple.

Elles font référence à quatre propositions de réponses possibles, qui peuvent être choisies à chacune des questions, mais qui ne sont pas répétées à chaque fois parmi les solutions proposées. Elles sont donc *implicites* et l'étudiant doit retenir que ces choix sont susceptibles d'être utilisés à chaque question.

Il s'agit des solutions suivantes :

- **6 ou AUCUNE** : aucune des solutions proposées n'est correcte.
- **7 ou TOUTES** : toutes les solutions proposées sont correctes.
- **8 ou MANQUE** : il est impossible de répondre parce que l'information (au moins une donnée) manque dans l'énoncé de la question (donc pas dans le cours ni dans la connaissance actuelle du problème).
- **9 ou ABSURDITÉ** : une absurdité dans l'énoncé rend toute la question sans objet (par exemple, une contre-vérité dans l'énoncé).

À titre d'exemple, voici quelques questions à choix multiple explicitant les différentes solutions générales implicites :

La capitale de la France est 1. Lille 2. Lyon 3. Paris	**3**
La capitale de l'Italie est 1. Berlin 2. Prague 3. Tokyo	**6**
La Grande-Bretagne comprend 1. L'Angleterre 2. L'Écosse 3. Le Pays de Galles	**7**
Quel âge Rimbaud avait-il ? 1. 2 ans 2. 10 ans 3. 20 ans	**8**
En quelle année Jules César a-t-il rencontré Napoléon ? 1. 1850 2. 1915 3. 1945	**9**

Attention ! La réponse 9 a priorité sur les autres solutions générales 6, 7 et 8 et évidemment sur les réponses dactylographiées 1, 2, 3...

Les degrés de certitude (DC)

Les degrés de certitude sont choisis par l'étudiant pour évaluer le niveau de connaissance de la réponse donnée.

Il s'agit d'une échelle en six points pour lesquels une certaine note est attribuée ou retranchée en fonction de l'exactitude de la réponse donnée.

Le barème des points a été calculé de manière à ce que les étudiants qui s'auto-évaluent bien, c'est-à-dire qui sont réalistes (ni sur-estimation, ni sous-estimation) par rapport à leur performance et qui disent la vérité, gagnent le plus de points. Ce barème est basé sur la théorie des décisions approuvées scientifiquement (Leclercq, 1967).

Si vous considérez que la réponse a une probabilité d'être correcte comprise entre…	Choisissez le degré de certitude…	Vous obtiendrez les points suivants en cas de réponse…	
		…Correcte	…Incorrecte
…0 % et 25 %	**0**	+13	+4
…25 % et 50 %	**1**	+16	+3
…50 % et 70 %	**2**	+17	+2
…70 % et 85 %	**3**	+18	0
…85 % et 95 %	**4**	+19	-6
…95 % et 100 %	**5**	+20	-20

Différents avantages et enjeux découlent de l'utilisation des degrés de certitude :

- L'ignorance est une situation normale de la vie et prendre conscience de notre degré d'incompétence aide à augmenter le niveau de compétence.
- L'auto-évaluation s'apprend par l'expérience personnelle.
- L'ignorance reconnue n'est pas dangereuse et il vaut mieux reconnaître les limites de ses compétences. L'ignorance dissimulée est quant à elle dangereuse.
- La production de jugement est un des niveaux d'objectifs les plus élevés.
- La connaissance n'est pas affaire de tout ou rien.
- Le doute est le moteur même de la connaissance.

Pourquoi nos procédures qualité ne prévoient-elles qu'une seule réponse correcte par question ?

Il est vrai qu'actuellement, nos procédures qualité prévoient une seule bonne réponse ainsi que l'utilisation des solutions générales implicites (notamment « *toutes les réponses sont correctes* » et « *aucune des réponses n'est correcte* »).

Les raisons de ce choix n'émanent pas de contraintes techniques mais sont avant tout **scientifiques et docimologiques**.

En effet, fournir des questions à choix multiple pouvant comprendre plusieurs bonnes réponses diminue fortement la qualité des tests. Cela revient à proposer des vrai-faux déguisés puisque l'étudiant doit s'interroger sur le caractère pertinent ou non de chacune des propositions.

Or, les docimologues savent que les vrai-faux sont bien moins performants que les questions à choix multiple. Dans son cours « *Évaluation des apprentissages* », Jean-Luc Gilles relève six raisons de ne pas utiliser de vrai-faux :

- Ils offrent peu de possibilités d'exploitation des erreurs. Il en découle une pauvreté des feedbacks potentiels.
- Ils offrent une fidélité théorique inférieure aux QCM.
- À fidélité théorique équivalente, il faudrait plus de temps pour répondre à un questionnaire vrai-faux qu'à des QCM.
- Ils surestiment les compétences des étudiants, puisque sans connaître la matière, on a une chance sur deux d'obtenir la bonne réponse.
- Ils manquent d'authenticité. De par leur nature dichotomique, les questions vrai-faux ne permettent en général pas de refléter des situations complexes « *authentiques* », comme celles que l'on peut rencontrer dans un contexte professionnel.
- À l'analyse, on remarque que, souvent, les vrai-faux évaluent essentiellement la connaissance et plus rarement les niveaux taxonomiques plus élevés.

En ce qui concerne plus particulièrement des questions à choix multiple où plusieurs réponses correctes sont possibles, nous pouvons citer, par expérience, deux autres inconvénients :

- Les enseignants ont tendance, dans ce cas de figure, à transgresser une des règles de rédaction de questions à choix multiple qui est que la question doit poursuivre un seul objectif. La plupart du temps, les questions à plusieurs réponses possibles partent d'une amorce très générale comme « *le cœur…* » et proposent des solutions évaluant chacune un objectif différent :

« *...est un muscle* », « *...bat à 65 pulsations chez un individu normal* », « *...a comme fonction de réguler la circulation du sang* »... Cette question, dans le cas mentionné, couvre respectivement les objectifs de connaissance de « *la constitution du cœur* », « *ses paramètres de fonctionnement habituels* », « *sa fonction* », soit autant d'objectifs différents.

- Ce type de question favorise les questions de détails, ce qui diminue la validité du test.

Rédaction des questions

Suite à l'élaboration de la table de spécification en lien avec la matière que l'on souhaite évaluer, il incombe au responsable d'évaluation de débuter la rédaction de ses questions à choix multiple.

Cette étape de création de questions permettra non seulement de construire une **banque de questions durable et stable** mais également d'avoir des items qui évalueront avec validité au moins un des objectifs d'apprentissage les plus importants référencés dans la table de spécification.

À ce niveau, nous proposons donc différents critères d'analyse formelle s'appliquant à la construction des questions à choix multiple :

- L'amorce pose-t-elle une question directe ou cerne-t-elle bien un problème spécifique ?
- L'item porte-t-il bien sur un corpus théorique plus large qu'un simple exemple directement issu des notes de cours ?
- Le vocabulaire et la structure de la phrase sont-ils simples à comprendre ?
- Chacun des distracteurs est-il plausible ?
- Les distracteurs et la réponse correcte sont-ils formellement équivalents ?
- Si possible, les distracteurs représentent-ils une erreur ou une incompréhension classique de la part des étudiants ?
- La réponse correcte est-elle indépendante des distracteurs ou d'autres items du questionnaire ?
- Existe-t-il une et une seule bonne réponse ?

En complément de ces quelques points à vérifier lorsque l'on rédige une question à choix multiple, il faut également tenir compte de toute une série de règles de rédaction, qui permettent d'avoir des questions de qualité d'un point de vue docimologique. Les 20 premières règles ont été développées par le Professeur D. Leclercq (1986, *La conception des questions à choix multiple.* Labor, pp. 85-107) et portent sur différentes thématiques, que nous allons décrire dans les lignes qui vont suivre. Les 8 suivantes sont issues directement de l'expertise du SMART.

1. Règles de rédaction concernant l'adéquation aux objectifs

R1. Respecter l'objectif :

Il convient de n'utiliser la QCM que si elle est le type de question le plus approprié pour mesurer ce que l'on désire évaluer.

R2. Coller à l'objectif :

La QCM doit correspondre à l'objectif visé, au comportement à évaluer.

R3. Ne pas perturber les apprentissages :

La QCM ne doit pas perturber les apprentissages. Il faut éviter les distracteurs pouvant fixer une erreur dans l'esprit de l'étudiant.

2. Règles de rédaction concernant la valeur diagnostique de la réponse

R4. Révéler le processus mental :

La QCM doit renseigner l'évaluateur sur le processus mental utilisé par l'étudiant.

R5. Indiquer l'erreur commise :

Les distracteurs doivent indiquer le type d'erreur commise ou le cheminement incorrect suivi par l'étudiant.

R6. Préciser sur quelle partie de l'énoncé porte la question :

Pour éviter un diagnostic erroné, on doit préciser sur quelle partie de l'énoncé porte la question (par exemple, en soulignant cette partie).

3. Règles de rédaction sur la forme

R7. Respecter la consigne :

La question doit être compatible avec la consigne.

R8. Proposer des phrases syntaxiquement correctes :

Les solutions doivent être en accord grammatical avec l'amorce.

R9. Éviter les termes vagues :

On n'utilise pas de termes vagues dans l'énoncé.

R10. Éviter les négations :

On évite les formes négatives (syntaxiques et sémantiques) et *a fortiori*, on proscrit leur accumulation. La formulation affirmative est préférable.

R11. Séparer informations et questions :

La question et les informations ne doivent pas être entremêlées.

R12. Regrouper dans l'amorce les éléments communs aux solutions proposées :

On fait remonter dans l'amorce et/ou on groupe à la fin de la question (en-dessous des solutions) les éléments communs aux solutions proposées.

4. Règles de rédaction des solutions proposées

R13. Indépendance syntaxique des solutions :

On n'utilise pas, par exemple, des expressions telles que « *au contraire* », « *en plus* », etc. au début des solutions proposées car ces expressions lient les solutions entre elles.

R14. Indépendance sémantique des solutions :

Les solutions proposées doivent être sémantiquement indépendantes les unes des autres. Deux solutions ne peuvent être emboîtées.

R15. Égalité des mots communs à la solution et à l'amorce :

On évite les mots communs entre l'amorce et les solutions ou on fait en sorte que chaque solution possède ces mêmes mots communs avec l'amorce.

R16. Égalité de vraisemblance des solutions :

On veille à une même vraisemblance des solutions proposées.

R17. Même longueur pour toutes les solutions :

La solution correcte ne doit pas être (systématiquement) plus longue que les autres. Les solutions doivent avoir une longueur équivalente.

R18. Même complexité de toutes les solutions :

La solution correcte ne doit pas apparaître comme plus complète que les autres.

R19. Même degré de généralité :

On privilégie un même niveau de généralité des indicateurs (temps, modificateurs…). *Tous, toujours, jamais, aucun* sont des termes absolus et catégoriques dont les étudiants se méfient. Ils préfèrent des solutions contenant les termes *certains, parfois, il peut arriver que*…

R20. Même degré de technicité :

On privilégie un même degré de technicité du vocabulaire utilisé dans toutes les solutions proposées. Les étudiants qui ne maîtrisent pas bien le contenu ont tendance à éviter les solutions comportant des termes techniques.

R21. Termes identiques pour une même idée :

On privilégiera autant que possible des termes identiques pour évoquer une même idée sur un même sujet. Lorsqu'une même idée est présente dans plusieurs solutions, on veillera donc à l'exprimer avec des termes identiques dans les différentes solutions. L'inconvénient d'une formulation différente réside dans le fait qu'elle pourrait engendrer des nuances non voulues par le rédacteur lui-même, perturbant ainsi le choix de réponse.

R22. Consensus sur le caractère correct ou incorrect des solutions :

Les propositions de solutions doivent permettre un consensus large sur leur caractère correct – lorsqu'il s'agit de la (des) réponse(s) correcte(s) – ou incorrect – lorsqu'il s'agit des réponses incorrectes.

R23. Équilibre entre les solutions positives et négatives :

Lorsque les solutions sont connotées positivement et négativement, il faut veiller à respecter un équilibre entre ces deux types de solutions, ceci afin de ne pas influencer la réponse des étudiants par un nombre plus important de solutions soit positives, soit négatives.

R24. Éviter de connoter les solutions de façon péjorative :

Il convient d'éviter de connoter les propositions de solutions. Il faut éviter d'utiliser des termes, des mots ou des expressions qui introduisent de fortes connotations qui, si elles sont négatives, auraient tendance à induire un rejet de la part de l'étudiant ou, si elles sont positives, pourraient artificiellement remporter l'adhésion de l'étudiant. On veillera à ce que cette question ne figure pas dans ce test, ou à supprimer cet indice.

5. Règles de cohérence dans le test

R25. Ordre logique :

Afin de ne pas influencer le choix d'une solution par sa position parmi l'ensemble des solutions proposées (par exemple, en mettant la solution correcte de manière régulière en position 1), il convient d'indiquer les solutions proposées dans un ordre croissant ou décroissant de grandeur. Cette règle s'applique aux solutions numériques mais vaut aussi pour l'ordre alphabétique ou logique. La cohérence (le choix d'un ordre croissant ou décroissant) dans l'application de cette règle au sein du test est à privilégier.

R26. Signes en toutes lettres :

Dans les questions où les propositions de solutions ne portent pas strictement sur des opérations numériques, il vaut mieux

éviter l'utilisation de signes (+, -, ×...). Certains signes pourraient être compris différemment d'une personne à l'autre. On privilégie donc l'écriture en toutes lettres de l'ensemble des idées évoquées dans les solutions proposées et ce, dans l'ensemble du test, afin d'éviter toute confusion. Ceci est également valable pour l'amorce.

R27. Uniformisation des ponctuations dans tout le test :
Il convient ici de mettre (ou de ne pas mettre le cas échéant) des majuscules en début et des points en fin de phrase pour toutes les solutions proposées. On recommande de procéder de la sorte pour les amorces également, afin que l'ensemble du test soit uniforme.

R28. Ne pas induire la réponse à une autre question du test :
Parfois, il arrive qu'une question mentionne des éléments permettant de déduire la solution à une autre question du test. Il convient donc d'être particulièrement vigilant lors de la création d'un questionnaire et de relire ce dernier pour vérifier qu'un souci de ce type n'est pas présent. Si l'on possède dans une banque de questions certaines d'entre elles apportant une aide pour la réponse à d'autres items, il est recommandé de signaler avec quelles autres questions elles ne doivent pas être présentées simultanément.

Nous pouvons aisément constater qu'une question à choix multiple présentant une bonne qualité docimologique n'est pas chose aisée à construire et requiert de garder à l'esprit les règles sus-mentionnées tout au long du processus de rédaction.

Suite à la phase de rédaction des questions, il peut être bénéfique de faire procéder à la relecture des questions par un expert, lequel pourra en valider la bonne construction, ce qui améliorera encore la qualité de l'évaluation mise en place. Le SMART peut vous aider à ce niveau (au niveau de la construction des questions, pas de la matière en elle-même).

Information et formation des étudiants

Afin que les performances des étudiants soient entachées du moins d'erreurs de mesure possible, il est nécessaire que suffisamment d'informations, voire de formations, soient fournies concernant la manière dont l'évaluation se déroulera.

Il est également nécessaire de communiquer les informations concernant le test et les conditions d'administration : quand et comment il sera administré, avec quelles modalités d'évaluation, expliquer sur quoi le test porte ainsi que ce dont il est composé.

Il est également recommandé de pouvoir laisser l'opportunité aux étudiants de s'entraîner à fournir les performances attendues afin, notamment, de pouvoir gérer leur anxiété et de leur donner l'occasion d'acquérir certaines compétences pour gérer les modalités de l'évaluation.

Indépendamment de la mise à disposition de certains tests d'entraînement, toute une série de conseils peut leur être communiquée préalablement à l'évaluation.

La stratégie de gestion du temps

- Commencer à travailler de manière productive tout de suite.
- Passer les questions pour lesquelles l'étudiant ne sait pas répondre, pour y revenir plus tard.
- Si le temps imparti n'est pas écoulé, l'utiliser pour relire les questions.

La mise en place d'une stratégie évitant les erreurs grossières

- Prêter attention aux consignes et à la question afin de déterminer clairement la nature de la tâche qui est attendue.
- Ne pas hésiter à demander des clarifications durant l'évaluation si c'est autorisé (mais pas sur le contenu).
- Vérifier sur le fond et sur la forme, chacune des réponses données.

L'adaptation de la stratégie au barème de cotation utilisé et l'utilisation, en cas de doute, d'un raisonnement déductif.

Lorsque des tests d'entraînement sont mis à disposition des étudiants, la meilleure façon de procéder est le recours au **test à blanc**.

Pour ce faire, il est conseillé d'organiser un test similaire à celui qui sera utilisé lors de l'évaluation et qui se structurera donc autour des mêmes :

- Objectifs d'apprentissage
- Types de documents (consigne, feuille de réponse…)
- Conditions de temps
- Modalités de questionnement
- Degrés de difficulté
- Procédures anti-fraude
- Conditions exogènes
- …

Le test en pratique

La commande des formuLOMs

Afin de pouvoir administrer l'épreuve que vous venez de créer, il est nécessaire de commander les **formulaires de Lecture Optique de Marques** (formuLOMs) *ad hoc* préalablement à l'épreuve. Cette commande s'effectue **exclusivement via notre site Web**, à l'adresse **http://smart.uliege.be** dans la rubrique *Enseignants — Commander des formuLOMs*.

Le SMART vous envoie alors les formuLOMs vierges par courrier (comptez un délai de 2 à 3 jours ouvrables pour leur réception). Pour chaque épreuve, un **formulaire de paramétrage** est fourni d'office (une ou plusieurs feuilles selon le nombre de questions) ainsi que les **feuilles de vérification**. Il s'agit des documents à remettre obligatoirement au SMART pour effectuer la correction de votre examen.

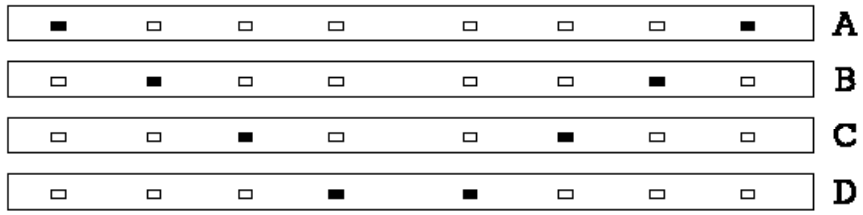
Il est également possible de commander des **feuilles de consignes** (concernant le **marquage**, les **degrés de certitude** et / ou l'utilisation des **solutions générales implicites**) en version papier pour les étudiants ou de les télécharger au format PDF, afin de les projeter durant l'épreuve par exemple.

Des **feuilles de justification** sont également disponibles à la commande. Ces feuilles permettent à l'étudiant de spécifier les raisons l'ayant conduit à choisir cette réponse. Suite à l'analyse de ces justifications par vos soins, il vous sera loisible de valider une réponse initialement considérée comme incorrecte pour ce seul étudiant et non pas pour l'ensemble du groupe.

Nous pouvons aussi vous faire parvenir des **bandelettes précodées** reprenant les cases à noircir sur les formuLOMs en fonction des formes de questionnaire proposées lors de l'épreuve.

Elles peuvent également être commandées (bandelettes autocollantes) ou **téléchargées afin d'être directement intégrées** dans le questionnaire à imprimer.

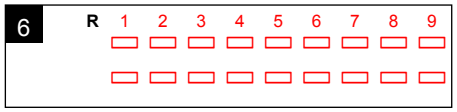
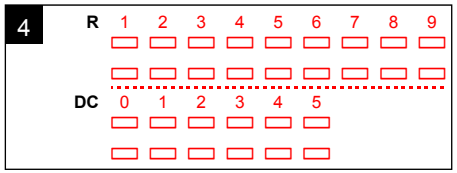
De la sorte, l'étudiant peut poser son formuLOM juste en-dessous de la bandelette et retranscrire les cases noires sans risque de se tromper.



Les modalités d'utilisation des différents types de formuLOMs

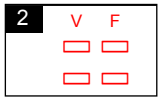
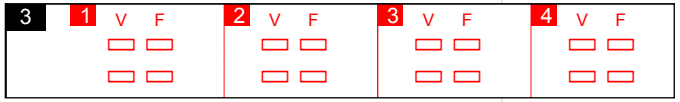
Question à choix multiple

Question composée d'une amorce qui pose le problème et de plusieurs solutions proposées parmi lesquelles se trouve une et une seule réponse correcte.

Modalité de questionnement	Type d'utilisation	Barème de correction	Consignes à associer
QCM « classique »  FormuLOMs disponibles en  ou  : • 45 questions *(recto)* • 102 questions *(recto-verso)*	• Jusqu'à 9 propositions possibles • Une seule réponse correcte	• Barème *for guessing* • Barème personnalisé	• Consignes de marquage pour les formuLOMs • Consignes concernant les solutions générales implicites
QCM avec degrés de certitude  FormuLOMs disponibles en  ou  : • 30 questions *(recto)* • 66 questions *(recto-verso)*	• Jusqu'à 9 propositions possibles • Une seule réponse correcte accompagnée d'un degré de certitude : 0, 1, 2, 3, 4 ou 5	• Barème utilisant les degrés de certitude	• Consignes de marquage pour les formuLOMs • Consignes concernant les solutions générales implicites • Consignes concernant les degrés de certitude

Vrai-Faux

En réponse à un énoncé, les seuls choix possibles pour l'étudiant sont Vrai ou Faux.

Modalité de questionnement	Type d'utilisation	Barème de correction	Consignes à associer
Vrai-Faux « classique »  FormuLOMs disponibles en 🇫🇷 ou 🇬🇧 : • 153 questions *(recto)*	• Une seule réponse correcte	• Barème *for guessing* • Barème personnalisé	• Consignes de marquage pour les formuLOMs
Vrai-Faux généralisé  FormuLOMs disponibles en 🇫🇷 ou 🇬🇧 : • 34 questions *(recto)* • 76 questions *(recto-verso)*	• Une amorce commune à quatre propositions pour lesquelles l'étudiant doit dire si elles sont vraies ou fausses	• Barème *for guessing* • Barème personnalisé	• Consignes de marquage pour les formuLOMs

Question à réponse ouverte moyenne

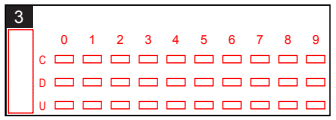
Question dont la réponse est constituée d'un mot, d'un chiffre ou d'une courte phrase.

Nécessite une intervention humaine dans la correction.

Modalité de questionnement	Type d'utilisation	Barème de correction	Consignes à associer
QROM  FormuLOMs disponibles en 🇫🇷 ou 🇬🇧 : • 18 questions d'une ligne *(recto)* • 40 questions d'une ligne *(recto-verso)* • 15 questions de trois lignes *(recto-verso)*	• Correction semi-automatisée : avant de nous fournir les formuLOMs, l'évaluateur devra juger du caractère correct ou incorrect de chaque réponse.	• Barème personnalisé	• Consignes de marquage pour les formuLOMs

Question à choix large

L'étudiant cherche parmi une liste conséquente de mots la réponse à sa question (il s'agit par exemple de l'index des concepts du cours), chaque terme étant référencé par un code qu'il devra reproduire sur son formuLOM.
Cette liste peut être utilisée notamment dans des textes lacunaires.

Modalité de questionnement	Type d'utilisation	Barème de correction	Consignes à associer
QCL à 3 chiffres : index entre 1 et 999  FormuLOMs disponibles en 🇫🇷 ou 🇬🇧 : • 30 questions *(recto)* • 66 questions *(recto-verso)*	• Un ou plusieurs code(s) correct(s)	• Barème personnalisé	• Consignes de marquage pour les formuLOMs QCL

La passation

Après avoir réalisé la conception de votre évaluation et avoir permis aux étudiants de s'entraîner, l'épreuve à proprement parler va pouvoir prendre place.

Il incombe au responsable de l'évaluation d'**établir un environnement sobre** lors de l'administration et d'adopter une attitude professionnelle. Il s'agit ici de ne pas bavarder avec les étudiants, de siffloter ou encore de rester juste derrière un étudiant pendant toute la durée de l'épreuve. Il convient également de répondre aux questions portant sur les tâches attendues lors de l'évaluation et d'adapter le test au temps disponible ainsi qu'aux étudiants présentant un handicap.

Il est également parfois nécessaire de prévoir des procédures anti-fraude via l'utilisation de questionnaires différents, où les questions ne sont pas présentées dans le même ordre, d'une forme à l'autre.

Le SMART autorise ici l'utilisation de **quatre formes de questionnaires**. Les questions pourront être mélangées **soit dans un ordre circulaire, soit dans un ordre aléatoire selon vos souhaits.** Nous vous conseillons de nous contacter pour plus de précisions à ce sujet.

Enfin, il est demandé aux étudiants d'utiliser **exclusivement un stylo à bille (Bic) noir ou bleu** pour noircir les cases de leur formulaire de réponse, le feutre ou le stylo-plume entraînant des *percements* sur le verso de la feuille et pouvant amener à la lecture de cases *initialement non noircies*. Le crayon est également proscrit.

Préalablement à sa réponse aux questions, l'étudiant devra noircir les cases correspondant à :

- **son matricule** (composé de 6 chiffres);
- **la forme de son questionnaire** (A, B, C ou D — dans le cadre supérieur du formuLOM).

En cas d'erreur, il lui est possible d'**utiliser la seconde ligne** de réponse et dans ce cas, seule cette réponse sera prise en considération. S'il le souhaite, il peut également utiliser un **ruban correcteur**.

Attention cependant à **ne pas retracer les cases rouges avec leur stylo à bille (Bic) noir ou bleu** car cela pourrait être considéré comme une réponse, notre lecteur optique ne lisant pas l'encre rouge (celle-ci permet seulement de *délimiter les zones possibles de réponses*, sans être lue).

La correction

Documents à remettre au SMART

Une fois votre épreuve construite puis administrée à vos étudiants, il vous faudra, bien évidemment, nous rapporter leurs **formuLOMs de réponse**, en vue de la correction de l'épreuve. Parallèlement à cela, il vous sera demandé de nous remettre **deux autres documents dûment complétés** (fournis lors de votre commande) :

- **La feuille de paramétrage**. Cette/ces feuille(s) (en fonction du nombre de questions) comprend/comprennent :
  * Les informations concernant le test (nom du professeur, intitulé de l'épreuve, nombre de questions).
  * L'utilisation ou non des SGI. Si oui, vous préciserez les **SGI utilisées dans tout le test** (6, 7, 8 et / ou 9).
  * L'utilisation ou non de **formes parallèles**. Si oui, vous indiquerez le nombre de questionnaires ainsi que le processus de mélange (circulaire ou aléatoire).
  * Pour **chacune des questions** composant le test : la *réponse correcte*, le *nombre de solutions proposées* (*hors solutions générales implicites)*; et, en option, le *chapitre / rubrique matière* (RM) et la *catégorie de performance* (CP) auxquels fait référence la question.
- La/les feuille(s) de vérification. Une feuille reprenant les réponses correctes pour chacune des formes de questionnaire est à remettre, même s'il n'y a qu'une seule forme. Il s'agit d'un contrôle qualité permettant de vérifier la correspondance des réponses. Ces feuilles ont un numéro de matricule particulier afin de les distinguer des feuilles des étudiants.

Informations à communiquer lors du dépôt des formuLOMs

En complément aux documents papier, différentes informations sont à nous communiquer lors du dépôt de vos formuLOMs :

- La **date de délibération**
- Le **nombre de décimales** pour les scores des étudiants
- L'**exportation** ou non des notes des étudiants sur myULiège
- La mise en ligne ou non des **feedbacks aux étudiants** et la date de leur disponibilité
- Le(s) **barème(s) de correction** utilisé(s)
- La **pondération** des questions (si différente de 1)
- Les intitulés de chapitres et de catégories de performance (optionnel)
- Toute remarque pertinente éventuelle concernant l'épreuve.

Barèmes de correction

Le choix du barème de cotation à utiliser doit être défini **préalablement à la correction**, pour que les scores calculés prennent celui-ci en considération. Les pondérations éventuellement utilisées pour certaines questions doivent également être connues à ce stade de la procédure, ce afin d'adapter le score obtenu à ces questions pondérées.

En ce qui concerne les barèmes de cotation des QCM, trois types sont généralement utilisés :

Le barème personnalisé

Il est **défini par le responsable de l'évaluation** et peut prendre n'importe quelle valeur, qu'il s'agisse de la réponse correcte, d'une omission ou d'une réponse incorrecte.

Ici, le responsable peut choisir d'attribuer (comme c'est souvent le cas) **un point en cas de bonne réponse, zéro en cas d'omission ou de réponse incorrecte**. Il peut également décider de retirer des points en cas de réponse incorrecte et/ou en cas d'omission.

Cependant, cette façon d'attribuer une note à la performance de l'étudiant peut ne pas prendre en compte, à sa juste valeur tout du moins, l'effet dû à une réponse donnée au hasard, ni valoriser le fait que l'étudiant reconnaisse son ignorance et/ou sanctionner le fait qu'il pense connaître une matière en réalité non-acquise. Ce biais peut toutefois être effacé en utilisant l'un des deux barèmes ci-après.

Le barème utilisant la correction *for guessing*

Comme son nom l'indique, ce barème prend en considération la probabilité de donner une réponse correcte à la question sans réellement en connaître la réponse. Dans le cas d'une question vrai-faux, l'étudiant **a 50 % de chance** de répondre correctement à la question, même s'il n'a aucune connaissance de cette réponse et donc, **50 % de chance de réussir l'épreuve**.

C'est pour pallier à cet inconvénient que la correction *for guessing* a été développée. Elle offre un point pour toute réponse correcte, mais retire également un certain nombre de points en cas de mauvaise réponse, poussant donc l'étudiant à se montrer prudent et à ne pas répondre au hasard lorsque la solution correcte lui est inconnue.

Le nombre de points retranchés dépend ici du nombre de solutions proposées (NSP), la formule étant la suivante en cas de réponse incorrecte : *-1 ÷ (NSP - 1)*.

À titre informatif, vous retrouverez ci-dessous les points retranchés en fonction du nombre de propositions disponibles :

Nombre de solutions proposées	Point retiré en cas de réponse incorrecte
2	-1
3	-0,50 (-1/2)
4	-0,33 (-1/3)
5	-0,25 (-1/4)

Comme on peut aisément le constater, les points retirés vont décroissant en fonction de l'augmentation du nombre de propositions. En effet, plus le choix de réponse est élevé, moins la chance de répondre correctement au hasard est élevée et donc, la « *punition* » pour réponse au hasard va diminuer pour s'adapter à ce fait.

En cas d'omission, l'étudiant ne se verra pas pénalisé : il ne gagnera pas de point mais ne s'en verra pas retirer non plus.

Le barème utilisant les degrés de certitude

Le dernier type de barème de correction que l'on peut utiliser dans le cas d'évaluations standardisées est celui utilisant les degrés de certitude. Dans cette configuration, en plus de demander à l'étudiant la réponse qu'il pense être correcte, on lui demande d'y associer le niveau de certitude avec lequel il pense que ce choix est véritablement la réponse attendue.

Comme dans le cas du barème *for guessing*, des points sont retranchés en cas de mauvaise réponse. Toutefois, ce n'est le cas que lorsque la certitude avancée par l'étudiant est élevée. Il s'agit ici **de pénaliser un étudiant pour une ignorance non-connue**.

Par contre, s'il répond de façon incorrecte en associant une certitude nulle, il fait preuve d'une reconnaissance de son ignorance et donc, se verra tout de même attribuer une note de 4 / 20.

Le tableau ci-contre reprend les notes obtenues en fonction du niveau de certitude et du caractère correct ou incorrect de la réponse donnée.

Ce barème de correction permet donc **à ceux qui s'évaluent bien**, qui sont réalistes sur leur niveau de connaissance, **de gagner plus de points** que lors de l'application d'un barème

Si vous considérez que la réponse a une probabilité d'être correcte comprise entre…	Écrivez…	Vous obtiendrez les points suivants en cas de réponse	
		Correcte	Incorrecte
…0 % et 25 %	**0**	+13	+4
…25 % et 50 %	**1**	+16	+3
…50 % et 70 %	**2**	+17	+2
…70 % et 85 %	**3**	+18	0
…85 % et 95 %	**4**	+19	-6
…95 % et 100 %	**5**	+20	-20

correctif tenant compte des probabilités d'avoir la réponse correcte en la devinant. De plus, **dire la vérité est la stratégie qui rapporte le plus de points**.

En plus de cette cotation impliquant de la part de l'étudiant de s'**auto-évaluer**, le responsable d'évaluation peut encore modifier la cote finale en fixant un **niveau de sévérité** particulier.

Comme nous l'avons vu, les cotes attribuées à chaque question sont sur 20, et le score global est lui aussi ramené sur 20.

Toutefois, il peut arriver qu'un responsable d'évaluation souhaite être moins sévère. Une procédure régulièrement utilisée est de transformer la note en un score sur 18, ou 16, et ramené ensuite sur 20. En guise d'exemple plus concret, prenons un étudiant qui a obtenu un 16 / 20. Il obtiendra un 17,7 / 20 en sévérité 18 : $(16 \div 18) \times 20$ ou un 20 / 20 en sévérité 16 : $(16 \div 16) \times 20$.

Utilisation de niveaux de sévérités différents plutôt qu'un ajout de points

Le niveau de sévérité est le concept mathématique permettant d'adapter la note obtenue à une épreuve en fonction du niveau d'excellence souhaité.

Il s'agit, de façon concrète, d'adapter une note initialement calculée sur 20, au niveau d'exigence souhaité (plus ou moins sévère). La plupart du temps, celui-ci est soit de 16, 18 ou 20.

De manière habituelle, **une sévérité de 18 est utilisée**, mais pour des matières impliquant par exemple de grandes responsabilités (telles que la chirurgie, l'aviation…), une sévérité 20 pourra être employée car les compétences à évaluer doivent être connues à la perfection par les étudiants et l'erreur n'est pas permise (Leclercq, D. (1998). *Pour une pédagogie universitaire de qualité*. Mardaga).

Cette façon de procéder est plus adéquate que celle consistant à ajouter X point(s) à tous les candidats pour différentes raisons, que nous allons décrire ci-dessous.

Explication théorique

Pour passer d'une notation classique à une notation conventionnelle, incluant la notion de sévérité, la formule à utiliser est la suivante :

$$\text{Note conventionnelle} : \frac{\text{Score}_{20} \times 20}{\text{Sévérité}}$$

À l'aide d'un exemple, transformons les notes fictives obtenues par des candidats en sévérité 16, 18 et 20 dans le tableau ci-contre.

	Sévérité 20	Sévérité 18	Sévérité 16	Ajouter 2 points
Candidat 1	10,45	11,611	13,063	12,45
Candidat 2	7,44	8,267	9,3	9,44
Candidat 3	6,89	7,656	8,613	8,89
Candidat 4	13,11	14,567	16,388	15,11
Candidat 5	5,44	6,044	6,8	7,44
Candidat 6	11,67	12,967	14,588	13,67
Candidat 7	15,78	17,533	19,725	17,78
Candidat 8	4,89	5,433	6,113	6,89
Candidat 9	8,11	9,011	10,138	10,11
Candidat 10	9,78	10,867	12,225	11,78
Candidat 11	13	14,444	16,25	15
Candidat 12	2,78	3,089	3,475	4,78
Moyenne	**9,112**	**10,124**	**11,390**	**11,112**
Écart-type	**3,867**	**4,297**	**4,834**	**3,867**

En cas de modification de la sévérité, on peut constater que, comme attendu, les scores et la moyenne au test augmentent si cette dernière diminue. L'écart-type associé à ces données se voit aussi augmenter au plus la sévérité diminue.

Par contre, si nous avions ajouté 2 points à tous les étudiants, nous pourrions aisément observer que seule la moyenne a augmenté de deux points, l'écart-type restant quant à lui inchangé.

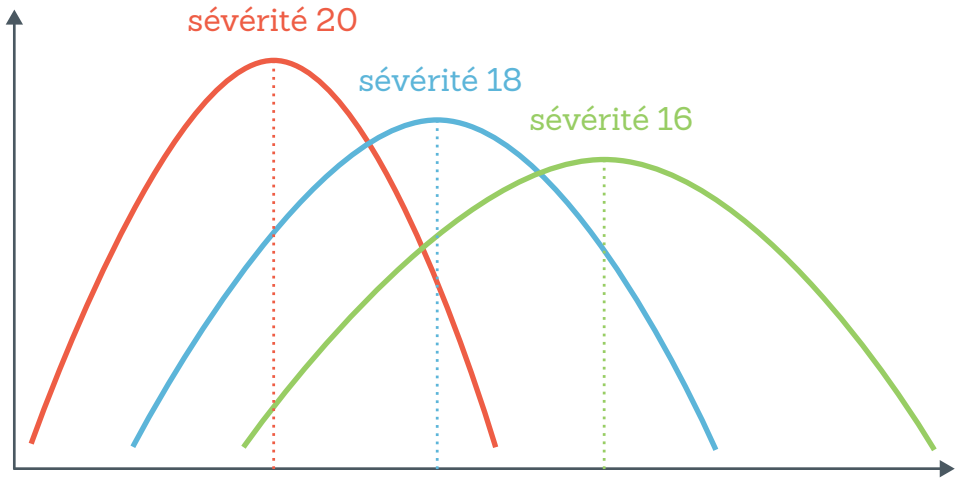
Pour rappel, l'écart-type sert à mesurer la dispersion d'un ensemble de données. Par exemple, la répartition des notes à une évaluation. Dans ce cas, plus l'écart-type est faible, plus le groupe évalué est considéré comme étant homogène, les notes restant fort proches les unes des autres. À l'inverse, un écart-type important signale que les étudiants ont des notes fort différentes les unes des autres, couvrant un éventail plus large du spectre des notes que l'on peut obtenir (par ex. de 0 à 20).

Dans le cas qui nous occupe ici, utiliser des niveaux de sévérités différents nous permet d'adapter la note des étudiants, mais de prendre également en compte la disparité des notes, leur variabilité, proportionnellement à leur variabilité initiale. À l'inverse, ajouter X points aux étudiants va en effet augmenter la moyenne du groupe, mais l'écart-type restera inchangé.

De la sorte, nous obtenons une mesure plus fidèle aux notes de départ. De plus, les étudiants ayant initialement mieux réussi, et donc ayant le plus de points, auront un gain de points proportionnellement plus important que ceux ayant moins bien réussi le test.

Par exemple, un étudiant qui a obtenu un 4,89/20 obtiendra un 6,113/20 en sévérité 16, mais un étudiant ayant obtenu une note plus élevée, un 13 par exemple, obtiendra quant à lui une note de 16,25/20 en sévérité 16. On voit ainsi que le gain obtenu suite au changement de sévérité dépend de la note initialement obtenue par l'étudiant.

Par conséquent, la distribution des notes va se déplacer vers la droite et va voir son amplitude augmenter au plus la sévérité diminue, contrairement au simple ajout de points à l'ensemble du groupe évalué.



Le fait de changer la sévérité utilisée possède donc trois avantages par rapport au simple fait d'ajouter des points aux candidats :

- Le **gain de points** relatif à l'utilisation d'une sévérité donnée va **dépendre de la note initiale** de l'étudiant. Un étudiant ayant moins bien réussi obtiendra moins de points supplémentaires que quelqu'un ayant mieux réussi.
- En utilisant cette méthode, l'écart-type des données va différer d'une sévérité à l'autre, avec un écart-type d'autant plus important que la sévérité est basse, entraînant de la sorte une augmentation de l'amplitude des notes et une **mesure plus fidèle à celle de départ**.
- Les étudiants qui avaient obtenu un score de 0 restent à 0.

Correction automatisée et contrôles qualité

Cette étape de correction est la partie la plus technique de l'ensemble du cycle que nous décrivons. Elle est composée de 3 phases essentielles :

1. L'évaluation de la production de l'étudiant ;
2. La réalisation de contrôles qualité sur la production et/ou les items de l'épreuve ;
3. Les modifications éventuelles consécutives aux contrôles qualité.

En ce qui concerne la correction des questions à choix multiple, sujet de cette partie du guide méthodologique, la correction se fait donc de manière automatisée, grâce à l'utilisation de **formulaires de lecture optique de marques** (*formuLOMs*). Cette façon de procéder permet de gagner en qualité et en objectivité, l'erreur humaine étant de la sorte minimisée, ainsi qu'en temps de correction, le logiciel utilisé étant nettement plus rapide qu'une correction manuscrite.

Lectures des formuLOMs de réponses

Une fois en notre possession, les formuLOMs de réponse vont être lus grâce à un *lecteur optique de marques*. Ils seront tout d'abord **lus à deux reprises**, avec des intensités de lecture différentes, permettant ainsi de **repérer d'éventuelles incohérences de lecture** (percements de l'encre à travers le papier, case mal noircie…). Ils seront ensuite **scannés** afin de pouvoir **prendre des actions sur les données lues**, sans avoir à effectuer des modifications directement sur les formuLOMs.

À ce stade, de nombreux **contrôle de la qualité** sont déjà effectués et plusieurs types d'erreurs sont détectés :

- Doublon de matricule lorsque deux étudiants ont malencontreusement codé le même identifiant.
- Incohérences de lecture et choix de la réponse souhaitée par l'étudiant.
- Réponses pour lesquelles l'étudiant a coché plusieurs cases sur la même ligne : suppression des réponses à ces questions.
- Erreur de codage de matricule : double coche ou oubli d'une ligne…
- Erreur/absence de codage de la forme du questionnaire. Dans ce cas, nous vous envoyons la liste du(des) formuLOM(s) incriminé(s) et vous demandons les actions que vous souhaitez prendre. Toutes les actions prises sur les données sont alors consignées dans les journaux de notre programme, permettant ainsi une **traçabilité maximale**.

Traitement des données

Une fois le fichier de données issues de la lecture des formuLOMs contrôlé et enregistré, la correction à proprement parler va pouvoir prendre place.

C'est ici que nous allons importer votre **feuille de paramétrage,** qui nous indiquera, pour chacune des questions, les réponses correctes, le nombre de solutions proposées ainsi que les chapitres et catégories de performance éventuels.

Nous importons alors **le fichier de données contenant les réponses des étudiants**. À ce stade, différentes erreurs peuvent également être détectées :

- Les étudiants n'existant pas dans le fichier signalétique : soit pour cause de codage d'un matricule inexistant, auquel cas nous lui réattribuons le sien ; soit car il n'est effectivement pas présent dans notre base de données, auquel cas nous l'ajoutons.
- Les étudiants étant inscrits dans deux sections/années différentes, que nous réinsérons dans celle qui convient pour le test traité.
- La présence de réponses dites « absurdes », n'étant pas possibles dans les choix de réponses.

Cela est généralement le signe d'une simple erreur de codage mais aussi :

* D'un percement de l'encre tellement important que notre double lecture ne l'a pas repéré et que la case percée a été considérée comme la réponse de l'étudiant. Dans ce cas, nous repassons l'ensemble de la feuille en revue et modifions les données si nécessaire.
* De l'utilisation de la deuxième ligne de réponse pour indiquer les degrés de certitude. C'est un fait assez exceptionnel mais qui s'est déjà produit, notamment pour les étudiants *Erasmus*, peu au fait de nos procédures.
* De l'utilisation de certaines solutions générales implicites (surtout les 6 et 7) par les étudiants, alors qu'elles n'étaient pas d'application. Dans ce cas, si une même question en montre beaucoup, nous vous le mentionnons afin de voir s'il n'y a pas quelque chose dans l'énoncé qui aurait amené les étudiants à répondre de la sorte et, le cas échéant, y remédier.

Ensuite, nous vérifions que le débrouillage indiqué (position de la 1ère question dans les formes parallèles) est bien correct et que le mélange de questions indiqué correspond aux réponses données sur les différentes feuilles de vérification que vous nous avez remises.

Enfin, avant de **générer les fichiers de résultats**, nous vérifions encore deux points essentiels :

- La similarité des moyennes obtenues entre les différentes formes du questionnaire;
- L'appartenance des étudiants à la bonne section. C'est le cas lorsqu'un étudiant a codé un matricule existant mais d'un étudiant d'une autre section. Nous le détectons ici et allons vérifier s'il s'agit bien de lui ou, dans le cas contraire, allons modifier ses données personnelles.

Afin d'assurer une **traçabilité** tout au long de la procédure, **une fiche de suivi** répertoriant toutes les actions effectuées sur les données des étudiants accompagnera l'épreuve.

Contrôles qualité par question : le *coefficient de corrélation point bisériale*

Une fois que les formuLOMs auront été lus et les données traitées, il sera possible d'effectuer une série de contrôles qualité sur les items de l'épreuve. Ces vérifications peuvent se faire grâce à l'analyse du *coefficient de corrélation point bisériale (r.bis)*.

Il s'agit d'un indice statistique édumétrique calculé pour chacune des propositions de chaque question et d'une corrélation linéaire entre le score global au test (variable *métrique*) et le choix pour chacune des propositions (variable *dichotomique* : choisie / pas choisie). En d'autres termes, cette statistique permet de vérifier si, en tendance, **les étudiants les meilleurs au test ont choisi la réponse correcte alors que cela ne serait pas le cas des étudiants les plus faibles**. À l'inverse, les distracteurs devraient présenter des taux de choix inférieurs pour les étudiants les plus forts au test, contrairement aux étudiants les plus faibles.

Le r.bis d'une proposition varie entre -1 et +1. Il est positif si la proposition est choisie, en moyenne, par les étudiants qui obtiennent un score total plus élevé au test et d'autant plus grand que la proposition est massivement choisie par les « *meilleurs* ». Un coefficient négatif correspond à la situation opposée.

Lorsqu'une QCM « *fonctionne* » bien, **on s'attend donc à un r.bis positif et suffisamment élevé pour la réponse correcte** et des r.bis négatifs ou proches de zéro pour les autres propositions.

Le r.bis de la réponse correcte peut être considéré comme satisfaisant s'il est supérieur à un seuil qui est fonction du nombre de questions de l'épreuve : il s'agit de l'inverse de la racine carrée du nombre de questions (n), soit $1 \div \sqrt{n}$.

Cela signifie que moins il y a de questions, plus le seuil que le r.bis de la réponse correcte devrait dépasser est élevé.

Le r.bis pour une proposition particulière fournit donc de précieuses informations sur **l'adéquation des distracteurs** des questions à choix multiple et permet d'apprécier dans quelle mesure ils sont ambigus ou discriminants.

Quand il s'agit de la proposition correcte, le r.bis permet de **vérifier si la question est réussie**, en moyenne, **par les étudiants qui ont un score global élevé au test**. Autrement dit, si ce sont les étudiants bien préparés qui ont répondu correctement et donc, **de voir si la question est discriminante**. Il renseigne aussi sur la cohérence de la question avec le reste du test.

L'examen du r.bis permet donc :

- de **détecter une incohérence** éventuelle entre le résultat à une question donnée et l'ensemble du test,
- d'**analyser la qualité des solutions proposées**.

Associé aux pourcentages de choix pour chaque proposition, il apporte une aide précieuse à la détection d'un problème de correction et/ou à la décision éventuelle de de poser différentes actions correctives sur le test, comme :

- Valoriser une/des autre(s) proposition(s), en plus de la réponse correcte. Dans ce cas, plusieurs réponses correctes coexistent pour un même *item*.
- Modifier la réponse correcte. C'est notamment le cas lorsque le responsable d'évaluation communique une « mauvaise » réponse correcte sur la feuille de paramétrage.
- Supprimer une question de l'épreuve si l'évaluateur la juge mauvaise *a posteriori* : mauvaise rédaction, ambiguïté dans l'énoncé…
- Valider une question du test au lieu de la supprimer, afin de ne pas pénaliser les étudiants qui avaient tout de même répondu correctement à la question.

Pour toutes les épreuves, avant d'envoyer les résultats, le SMART analyse les statistiques des questions ainsi que les statistiques descriptives (*moyenne, écart-type, graphique de fréquence*…) et envoie un mail personnalisé au responsable d'évaluation mentionnant les potentiels problèmes relatifs à son épreuve.

Envoi des résultats

Une fois le processus de correction terminé, nous vous faisons parvenir par e-mail, le fichier PDF de la première version des résultats de votre épreuve, contenant le détail :

- des scores des étudiants (score global sur 20, score par partie, score par catégorie de performance) ;
- de l'anayse de la qualité des questions à l'aide du *r.bis* ;
- des statistiques du groupe (moyenne, médiane, écart-type, courbes de fréquence…).

Nous rappelons également dans notre mail les paramètres de correction utilisés **ainsi que les questions pour lesquelles les r.bis semblent problématiques**.

Après analyse des résultats, vous avez la possibilité de prendre d'éventuelles décisions de rectification sur votre épreuve :

- suppression ou validation d'une question ;
- changement de réponse correcte ;
- valorisation d'une autre proposition ;
- changement de barème de correction.

Par retour de mail, il vous suffira de nous indiquer les modifications à réaliser et nous vous renverrons alors le fichier PDF de la version actualisée des résultats de votre épreuve.

Validation de l'épreuve

Une fois que vous aurez pu parcourir le document PDF et que les modifications éventuelles auront été effectuées, vous devrez nous faire part de votre décision quant à la validation de la version de votre épreuve qui convient.

Sans cet accord, aucune validation ne sera réalisée et dès lors, l'envoi du **fichier de résultats définitifs, sous format tableur** (Microsoft Excel) ainsi que la **mise en ligne des feedbacks aux étudiants** (si vous l'avez demandée) ne seront pas effectués.

Cette validation se réalise au travers de l'envoi d'un mail stipulant le **cours, le code test associé (AXXXXXX) et la version à valider**. Nous vous ferons alors parvenir le fichier tableur des scores et nous vous informerons de la création des feedbacks à la date demandée.

Exportation des notes via myULiège (disponible uniquement pour les enseignants de l'Université de Liège)

Si vous le souhaitez, il est possible, une fois que vous avez validé votre épreuve, d'aller **charger les notes de vos étudiant directement dans votre espace myULiège** et de les transférer vers la base de données *Pénélope*. De la sorte, les **erreurs d'encodage sont éliminées** et le **gain de temps est considérable**.

Les feedbacks aux étudiants

Une fois l'évaluation réalisée, corrigée et validée, le responsable d'évaluation peut s'il le souhaite **proposer un feedback aux étudiants**, disponible sur leur espace myULiège (pour les étudiants de l'Université de Liège) ou sur le site web du SMART (pour les étudiants extérieurs), à une date définie. Dans ce feedback apparaîtra toute une série d'éléments, comme **l'exactitude des réponses données** aux différentes questions, le **score par partie** s'il y en a, ou encore **le niveau de réalisme** si les degrés de certitude ont été utilisés.

Pour les enseignants de l'Université de Liège, vous pourrez, si vous le souhaitez, y adjoindre également le **questionnaire d'examen**, directement depuis votre espace myULiège et modifier la date de publication.

C'est ici qu'**intervient toute la richesse de la table de spécification**. Nous l'avons signalé, les questions doivent permettre d'évaluer des **objectifs d'apprentissage** définis comme un croisement dans la table de spécification. Par ce fait, nous pourrons rendre un **feedback sur la maîtrise d'un objectif d'apprentissage** donné, mais également sur la **maîtrise d'un chapitre** ou sur la **maîtrise d'une catégorie de performance particulière**.

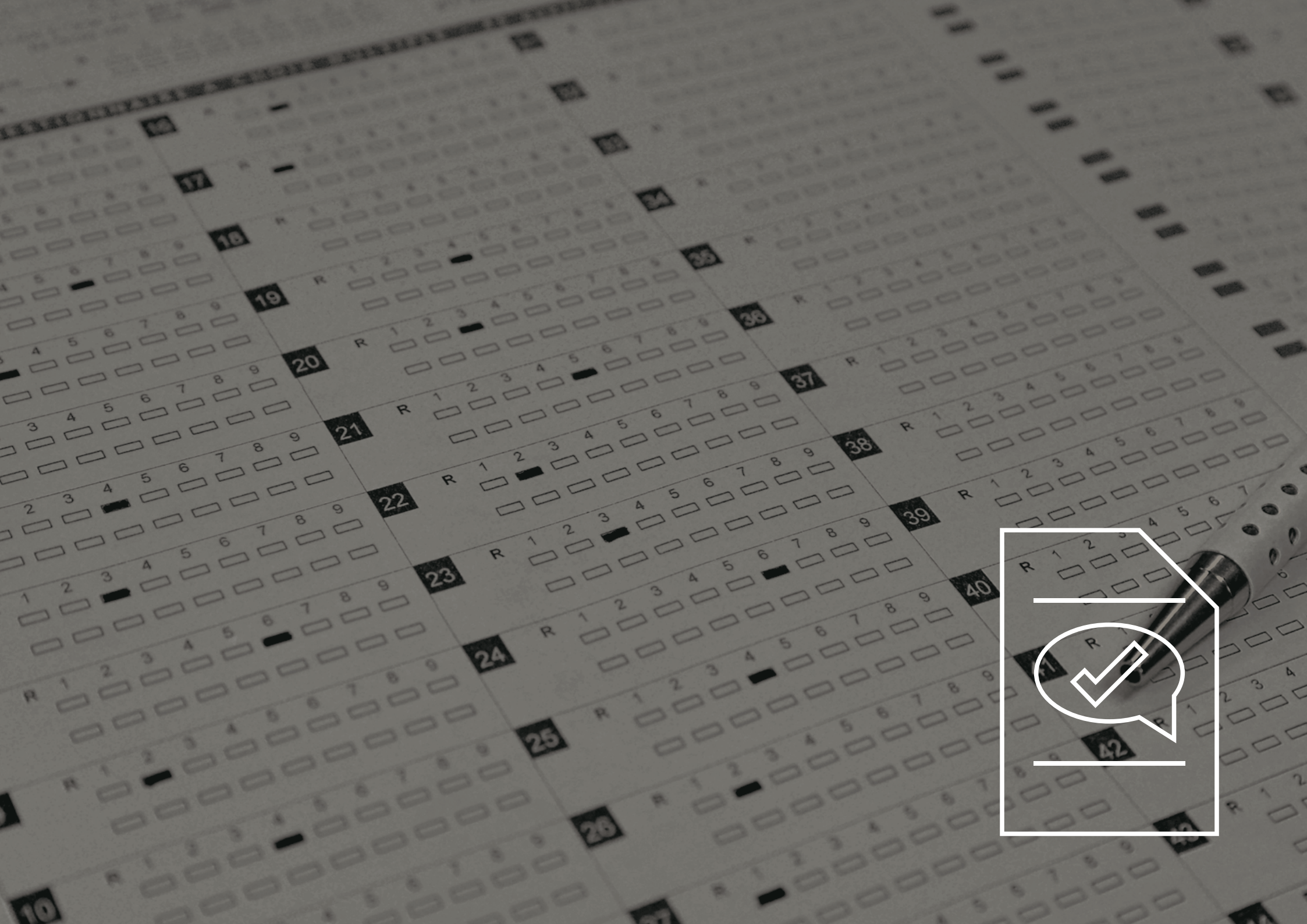
La table de spécification s'érige donc comme un outil de structuration du feedback qui offre aux étudiants une vision très diagnostique de leur performance. À noter qu'un feedback de ce type ne peut être délivré que s'il est géré de manière automatisée.

Il s'agit donc d'avoir ici une visée **formative**, même si l'évaluation était initialement à visée **certificative**.

Les feedbacks aux professeurs

Si vous ne souhaitez pas que vos étudiants aient accès à leur feedback en ligne, mais que vous organisez une **séance de consultation des copies**, nous pouvons également vous fournir un « *pack feedback* » reprenant l'ensemble des feedbacks de vos étudiants, au format PDF.

Vous aurez donc directement sous les yeux le **caractère correct ou non des réponses données par l'étudiant** (réorganisées dans l'ordre de la forme A du questionnaire) ainsi que la **note globale, par parties et par catégories de performance** (si utilisées). Si les degrés de certitude ont été utilisés, il vous sera également possible de voir **dans quelle mesure chaque étudiant s'auto-évalue bien** et, dans le cas contraire, s'il a tendance à la sur- ou à la sous-estimation de ses connaissances.



version : 2023-05-15

Photographies pp. 2, 4, 12, 23, 26, 29, 30 et 41 :
© Michel Houet, TILT-ULg
Photographies pp. 2, 30 (bandeaux supérieurs) et 61 :
© Damien Depluvrez, SMART – IFRES – Université de Liège

SMART — Système Méthodologique d'Aide à la Réalisation de Tests

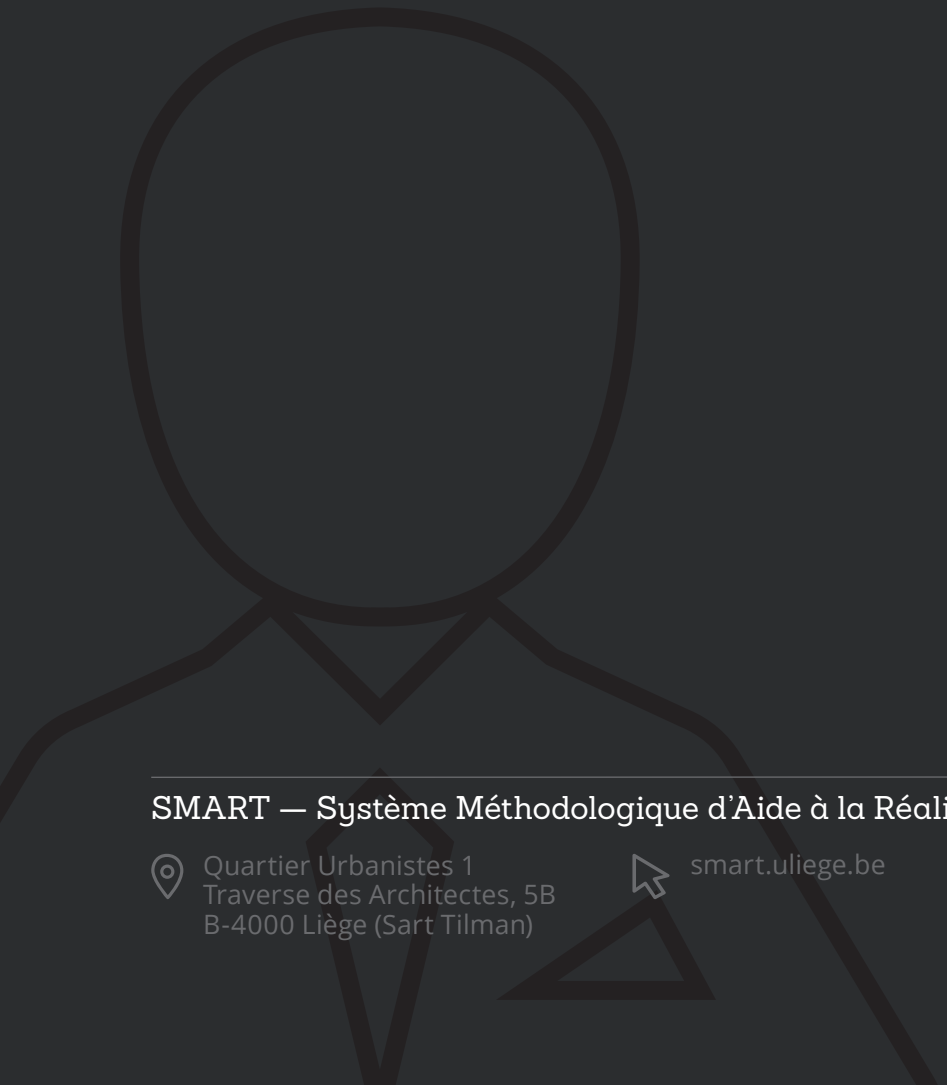
Quartier Urbanistes 1
Traverse des Architectes, 5B
B-4000 Liège (Sart Tilman)

smart.uliege.be

+32 4 366 2078

smart@uliege.be

LIÈGE université



SMART — Système Méthodologique d'Aide à la Réalisation de Tests

Quartier Urbanistes 1
Traverse des Architectes, 5B
B-4000 Liège (Sart Tilman)

smart.uliege.be

+32 4 366 2078

smart@uliege.be